2010 IEEE-RAS International Conference on Humanoid Robots
Nashville, TN, USA, December 6-8, 2010

Person Identification through Perceptual Fusion

E. Martinson, W. Lawson, J. G. Trafton

Abstract—When a robot interacts with an individual, it is important to know with whom it is interacting, either to avoid social faux pas or remember user preferences. Continuous person identification during normal interactions, however, is extremely challenging. A person is periodically speaking to the robot, while at the same time changing pose, looking in other directions, etc. In this paper, we address the problem of continuous person identification using both speech and face recognition. We demonstrate that both modalities can together produce a system that is superior at person identification than from using a single modality alone.

I. INTRODUCTION

Under near optimal conditions, people are adept at identifying others through voice or face even if they do not have a large amount of experience with them [1]. As should be expected, however, under less ideal circumstances, person identification degrades, visually [2] and aurally [3], yet people still recognize other people. Clearly, a combination of these cues (and others) is more effective under such situations than either speaker or face recognition alone [4]. Multiple cues are especially useful when one (but not both or all) cue is degraded in some way. Our goal in this paper is to provide humanoid robots with the ability to recognize people through multiple cues, fusing results from different perceptual domains to provide superior recognition.

One of the most interesting findings in human-robot interaction is that people have certain expectations of embodied agents [5]. These expectations can be set by the physical attributes of the robot (e.g., how the robot actually looks and sounds) [5-7]. Since humanoid robots have so many shared characteristics with people, it sets people's expectations very high. Thus, it is critical to be able to provide humanoid robots with several basic capabilities; robust person identification in a variety of different situations and circumstances is one of these key components of human-robot interaction.

In this paper, we present the results of work towards achieving our goal of natural person identification in human robot interaction. Our humanoid robot platform is the Mobile-Dexterous-Social (MDS) Robot[1] called Octavia. The MDS robot neck has 18 DoF for the head, neck and eyes allowing the robot to look at various locations in 3D space (pan, tilt). Perceptual inputs include two color video cameras, an SR3000 camera to provide depth information, and a 4-element microphone array. For this work, Octavia moves its head and torso to change its visual field of view in response to sound source localization [8]. Person identification is then performed using two separate systems: face recognition and speaker recognition. The two are fused together by a combination of decision-level and score-level fusion. Together with sound source localization, the fused recognition engine can dynamically track a multi-speaker interaction with the robot.

This paper is organized as follows: related work in face and speaker recognition is presented in the next section; section III describes the face recognition system, including results from a user study; section IV summarizes the auditory systems for speaker recognition and sound source localization; and section V describes the fused system and its advantages.

Fig1. Octavia is an MDS robot with a 4-element microphone array mounted on the body and cameras in the eves.

Manuscript received June 18, 2010. This work was partially supported by the Office of Naval Research under job order number N0001408WX30007 and 09-Y861 awarded to Greg Trafton.

E. Martinson is a post-doctoral fellow with the U.S. Naval Research Laboratory, Washington, DC 20375. (phone: 202-404-4948; e-mail: eric.martinson.ctr@nrl.navy.mil).

W. Lawson is also with the U.S. Naval Research Laboratory Washington, DC 20375. (e-mail: ed.lawson@nrl.navy.mil)

[1] http://robotic.media.mit.edu/projects/robots/mds/overview/overview.html

II. Related Work

Face recognition is a widely studied biometric that has attracted the attention of many researchers in different fields. Turk and Pentland [9] developed the well-known "Eigenfaces" approach for face recognition. Features are extracted using principle components analysis (PCA) on training images. A probe image is classified by comparing the Eigenvalues of the probe images against Eigenvalues of the gallery images. Others have explored Independent Components Analysis (ICA) [10] and Linear Discriminant Analysis (LDA) [11] for feature extraction. Recent work by Wright et al. [12] explored the use of sparse coding for face recognition. Features are extracted from a set of training images using l^1 minimization. The quality of the match is evaluated from reconstruction error.

Similar to face recognition, speaker recognition is the ability to identify a speaker from their voice. In robotics, speaker recognition has been most commonly tied to speaker position. By separating out different speech streams and localizing the speaker, the speaker is identified for purposes of interaction [13]. More recent work has focused on improving tracking of those moving speakers [14], but speaker recognition from voice is uncommon. An exception is work by Krsmanovic [15], who demonstrated speaker recognition on a robot using unique phrases for each speaker in conjunction with a Markov decision process. In contrast, this paper addresses the more general problem of text independent speaker recognition, where a user can say anything and be identified by their voice. A solution common to telephones and tele-conferencing applications are Gaussian mixture models [16]. More recent work, which may be of particular interest to robotics, are efforts to account for variable speaker positions and signal-noise ratios [17]. Without enough training data, new algorithms are necessary to avoid mismatched noise conditions, and lower precision.

The combined problem of face recognition and speaker identification is an emerging area for improving recognition rates. Palanivel and Yegnanaravana [18] use decision level fusion to integrate speech, face recognition and visual speech using a weighted confidence measure. This work uses a database created from news anchor footage to demonstrate an increase in overall recognition by combining modalities. We expand upon this concept by considering situations where either the face is not visible or the person is not speaking.

Outside of biometrics, there have been fusion efforts in recognizing toys that are partially occluded, visually and aurally [19]. This work, which fuses audio and visual coefficients in real-time using a neural network, shows a clear advantage in identifying objects hidden from the robots eyes or ears.

III. Robot Vision

We use the face recognition approach developed by Kamgar-Parsi et al. [20]. Our approach operates under the assumption that we are able to recognize certain people, i.e. those that Octavia "knows", while rejecting others as unfamiliar, much like humans. This approach is based on identifying and enclosing the region R_T in the human face space that belongs to the target person T.

Face recognition first requires a training phase before it can be used to identify a target person. When the system is tested, if a face is projected inside R_T it will be identified as the target person, otherwise rejected as being T. During training, however, the region R_T is identified with the help of a human critic. Suppose we have the image I_T of the target person T, and a large database, F, of facial images containing images $f_1,f_2,...,f_n$. The image I_T is morphed toward the image f_k (k is an element of n) until it becomes borderline acceptable, i.e., significantly different from I_T, yet still recognizable as T according to the human critic. Next, I_T is morphed even further toward f_k until it becomes borderline unacceptable, i.e. some resemblance to the target person, but not enough to be recognizable as that person. An example of how the region R_T in the human face space is enclosed is shown in Figure 2. In Figure 2, the second left picture is still recognizable as Jennifer Aniston (positive borderline exemplar, blue dots), while the next picture is not (negative borderline exemplar, red dots). Likewise, morphing Jennifer's image towards many other people will generate sufficient landmarks to identify and enclose the space belonging to Jennifer (shaded area).

Morphing percentages are established with the help of a human critic. Typically, a human critic would need to examine morphed images of the target person, I_T, toward only some 10 to 20 images in F to determine the average morphing percentages for the borderline acceptable and unacceptable exemplars. Applying those percentages and using images in F, the computer will then automatically generate and label a large training set. In practice, it is adequate if most of the generated exemplars are projected where intended, because an appropriate classifier will fit hyperplanes to the generated data.

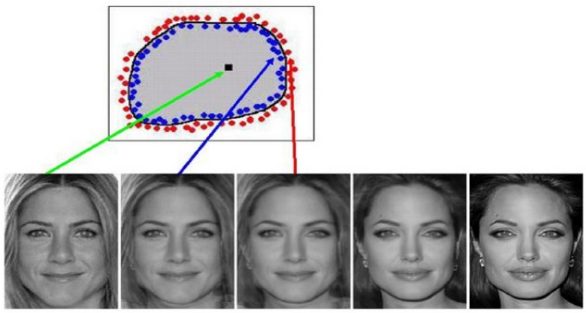

Fig 2. Jennifer Aniston, the leftmost picture, is morphed towards Angelina Jolie, the rightmost picture

The remainder of this section discusses: 1) the details of the image morphing procedure; 2) the training procedure, which uses the morphing results as part of the training process; and 3) the performance of the face recognition system on Octavia.

A. Image Morphing

To morph two faces, we warp using facial landmarks then cross-dissolve the warped images. We locate facial landmarks using Active Shape Models (ASM), a statistical model that captures shape. They fit a shape by iteratively moving local feature points towards an individual landmark, then fitting of all points using the statistical model. This process that repeats until ASM has converged.

We warp the eyes and nose region of the face using landmarks detected on the eyes, nose, and eyebrows. A Delaunay triangulation $T=<v,e>$ builds a mesh of triangles whose vertices (v) are facial landmarks connected by edges (e). To warp two faces (f_1 and f_2) using a morphing percentage m, we move the landmarks l_1 and l_2 to a new location $l_m = ml_1 + (1 - m)l_2$. We find the Delaunay triangulation T_m using l_m. The process of morphing involves moving the respective landmarks from l_1 and l_2 to their new location l_m. We warp each triangle by sampling pixel values using Barycentric coordinates. The morphed image I_M is the two warped faces combined using a percentage (p) of F_1 and (1-p) of F_2.

B. Training Classifiers

We build a dedicated neural network to recognize each individual that Octavia knows. To create these neural networks, we must begin by collecting a sufficient amount of training data. We train each neural network starting with a single prototypical face. We evaluate the resulting network; incrementally add faces that do not project inside of R_T. We repeat this process until the manifold for the subject established from the training data has been properly captured. In our experiments, we find that anywhere between 10-15 faces are sufficient to describe R_T.

We perform a similar incremental process to find the set of images to warp against. We begin with a small set of randomly selected images, and then incrementally add images of imposters if they project inside of R_T.

C. Online Recognition

We train Octavia to recognize the faces of people that are looking at her. That is, we focus mainly on frontal images of people. The first step in recognition is face detection, which we perform using the Viola-Jones face detector in OpenCV. Our face detector is tuned to recognize frontal faces within approximately 2m of the robot.

Next, pupils are located using an eye detector. The pupil detector looks for dark regions in the approximate area of the eyes of the subject. Detected eyes are moved automatically by the system to a canonical position to align faces. Finally, we process the aligned face to normalize intensity.

Recognition involves two steps. The first is to check the pose of the subject using a dedicated pose-network. The pose network was trained on a set of frontal and profile images. The purpose of the pose network is to further reduce the number of images processed to only those where the subject is looking at Octavia. The second step is to evaluate each face using the dedicated neural networks. We use a sliding window to evaluate the scores from each dedicated network over a three-frame window. The resulting score is compared to a threshold. If it exceeds this threshold, we say the person is recognized.

Figure 3 shows the ROC curves evaluating our approach. We evaluate our results using a set of individuals that Octavia knows. Each subject participated in two different sessions. The first session was used to train the network; the second session was used to test the network.

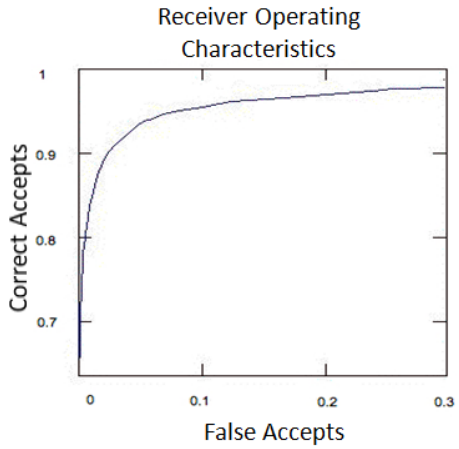

Fig 3. ROC curve of our approach to face recognition. The y-axis shows the hit rate, while the x-axis shows the false accept rate

IV. Robot Audition

Octavia is equipped with 4 lavalier type microphones: 2 on the body, and 2 on the backpack in the rear. Data acquisition is performed by a dedicated DMM-32X-AT PC104 card. This auditory information is currently used for two distinct purposes. The first is speaker recognition, which currently only uses feedback from a single microphone mounted on the left upper body. Samples are collected at 8192 Hz.

The second use of the microphone array in this work is speech localization. Speech localization is critical for fusion, because it augments face recognition very nicely. Without the means to rotate to a speaker, face recognition is limited to a very narrow field of view. By enabling robotic rotation towards a detected speech source, the applicability of face recognition system is expanded to more general human-robot interaction scenarios. In contrast to speaker recognition, speech localization utilizes all four channels.

A. Speaker Recognition

Speaker recognition on Octavia is based on Gaussian mixture models (GMM), as described by Quatiri [16] for use in speaker verification. Individuals are first recorded in a training session using the onboard microphone array. Then that training data is used to create a model for use in real-time recognition of that specific speaker.

To build a model of the speaker, recorded training data is processed to extract the first 10 mel-frequency cepstral coefficients (MFCC's) for each 10-msec frame. This produces a set of MFCC vectors. As the first MFCC is correlated to the energy of the audio segment, we apply a threshold to it for separating speech from ambient noise. This speech detection method assumes that all loud, wide spectrum sounds heard by the robot are actually speech, but is otherwise effective for removing quiet frames. A GMM with 50 components is then created from all remaining speech vectors using k-means. The resulting model consists of a centroid μ_i, a covariance matrix σ_i, and a prior probability p_i for each component i. Except to identify the presence of speech, the first two MFCC's are not used in model creation, leaving only 8 dimensions ($R = 8$).

Models are created both for the speaker and an "imposter" created from all speech frames for other speakers in the data set. Given a speaker/imposter pair $\{S_k, I_k\}$, which will henceforth be called a speaker verifier, the odds of an audio segment belonging to speaker k are calculated as follows. First, MFCC vectors x_m are extracted for the entire segment and quiet frames are removed. Next, for each vector, the probability of belonging to either model M is determined

$$p(\vec{x} \mid M) = \sum_{i=1}^{50} \frac{p_i}{\sqrt{(2\pi)^R |\sigma_i|}} e^{(\vec{x}-\vec{\mu})^T \sigma_i^{-1} (\vec{x}-\vec{\mu})}$$

To avoid zeros in subsequent calculations, it is assumed that each vector had to belong to either the speaker or the imposter, and that each were equally likely. So, given an audio stream segment α_t, which contains a set of speech frames m, the odds of a speaker being present at time t are:

$$O_k(t) = O_k(t-1) + \sum_m \log\left(\frac{p(\vec{x} \mid S_k)}{p(\vec{x}_m \mid I_k)}\right), \forall m \in \alpha_t$$

For use with continuous streams, O_k is restricted to the range [-15,15] preventing extremely large values. A decay function is also added to bring all speaker likelihoods back to neutral (i.e. 50% or O_k=0) when no one is talking. Whenever an audio segment contains no speech, O_k is reduced by -0.2*O_k across all speakers k.

B. Speaker Recognition Performance

Speaker recognition for human-robot interaction presents a different set of challenges from previous recognition work in telephone related environments. People move in relation to the microphone, resulting in a varying, and low, SNR, the microphone(s) on the robot are not ideal for speech, and robot ego-noise varies over time as different motors are engaged and fans turned on/off.

Therefore, to evaluate performance of a GMM based algorithm in a robotic scenario, a small user study was initiated. A total of 11 participants were involved in the study, but one was discarded for speaking too quietly for speech detection. The study was divided into two sections. First, data was collected from participants for building a speaker model during a training session. Then, participants were asked to speak on arbitrary topics from four different positions surrounding the robot to build test data set. During the training session, participants were asked to repeat a series of words and phrases, as well as speak freely for 20-sec on two subjects. The specific training protocol using for this work was based on the protocol used in creating the CSLU Speaker Recognition Corpus [21]. We, however, did not end up using the repeated words and utterances as part of our training data, because it lowered performance when recognizing speakers from free speech interaction. Furthermore, instead of contacting participants on different days, they were asked to repeat the protocol from two different locations relative to the array to include likely variations in speaking distance and incident angle.

All testing data was then broken into 2-sec fragments for evaluation. Fragments not containing speech were discarded, leaving approximately 32 fragments per speaker. Each fragment was then analyzed by sets of verifiers, ranging in size from 2-8 speakers. At least one of the speakers in the verification set was the actual speaker, and different imposter models were constructed for each set of speakers analyzed. A successful classification, or true positive, was where the correct speaker had the highest log odds likelihood, and that likelihood was greater than a minimum threshold (i.e. 10). A false positive occurred when a verifier for an incorrect speaker scored highest, and higher than the same minimum threshold. Fragments for whom all verifiers scored below the minimum threshold were considered unclassified.

Figure 4 plots the resulting precision and classification rates versus group size. With only two speakers in a set of verifiers, the precision, or rate of true positives to both true and false positives, averages 92%. Even the worst performing speaker model has an 85% precision rate. These

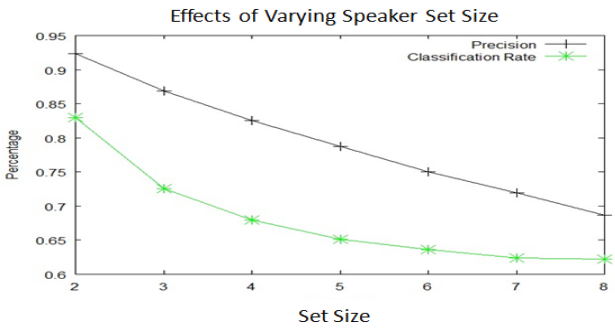

Fig 4. Precision and classification rate decrease with the number of speakers in the set.

rates drop rapidly with an increasing group size. By six speakers, the mean precision is only 75%, and the worst speaker model rates 50%. This, however, depends highly on the speakers in the set. One speaker in our test set had a significant spoken accent. That speaker's model achieved 100% precision.

The classification rate, or the number of samples generating likelihoods greater than the minimum threshold vs total number of samples, also dropped with the size of the verification set. With only 2 speakers in the set, 83% of the samples are classified, dropping to 64% by 6 speakers. Analyzing the minimum threshold, which was set experimentally prior to the trial, revealed that lower thresholds would raise the classification rate while dropping the precision across all group sizes.

C. Speech Localization

Speech localization on Octavia is determined with an auditory evidence grid [14], a time delay on arrival based localization method. Every half-second sample of audio data from a four microphone array is evaluated for the presence of speech using the MFCC-based method described previously in section IV.A. Those samples containing speech are then analyzed using a generalized cross correlation algorithm to estimate the energy from a hypothetical set of sound sources located 1-m out from the robot at 1-deg intervals in a 180-degree arc. The energy across all hypothetical sources, when normalized, is a spatially organized likelihood of sound source location that can be combined over time using log-odds notation. When one strong, or multiple weaker measurements from a single speaker creates enough evidence (e.g. a threshold), Octavia rotates to face the speech target.

As with speaker identification, samples not containing speech should lessen the likelihood of any speech source being present. This is performed by introducing neutral likelihoods, where all locations are equally likely, for all non-speech samples.

V. Integrated Robotic System

Speaker and face recognition have different strengths and weaknesses, but are based on similar classification models, making them ideal for fusion. Face recognition, for instance, requires that an individual is looking at the robot, and that lighting conditions not change dramatically. Speaker recognition requires people to speak and to speak sufficiently loud. Both systems enable robotic recognition of people some of the time, but neither is adequate for continuous recognition of an active participant. Combining the two classifiers together limits the times in which the robot is unable to identify the speaker.

In this section, the two recognition systems are fused together using a combination of decision-level and score-level fusion. Subsection A demonstrates performance advantages of the combined approach. In subsection B, sound source localization and robotic movement is added to the combined system to enable multi-speaker person identification.

A. Fused Recognition

Both speaker recognition and face recognition use a threshold based classification system, where a model is created and used to estimate the likelihood of that person being present or not. The only complication in combining the two systems is their asynchronous recognition rates. Face recognition is completed at ~5 frames per second, while speaker recognition is updated in 10-msec increments. To combine the systems, an arbitrary synchronization time of 0.5-sec was selected. Every half-second, face recognition is queried to retrieve the convolved response per speaker at time t, $\Phi_k(t)$. The convolution is a weighted average of the last 4 scores for each person in the watch list resulting in a likelihood with range [0,1]. For speaker recognition, the odds of each speaker having said anything are collected at time *t*. To give speaker recognition the same range as face recognition, a constant offset is added to $O_k(t)$ and the combined score is divided by 30 ($O_{max} - O_{min}$). The two systems are then fused by weighted summation giving equal weight to both face and speaker recognition:

$$\zeta_k(t) = \frac{\left(\Phi_k(t) + \left(O_k(t) - O_{\min}\right)/\left(O_{\max} - O_{\min}\right)\right)}{2}$$

This score level fusion, however, does not make sense when there is reliable feedback is missing from either face recognition or speaker recognition. If no one is speaking, as determined by the first MFCC value (see Section III.A), then only face recognition means anything. Similarly, if the speaker is looking to the side, or otherwise not visible to the robot's cameras, then speaker recognition results should be used alone. Therefore, to boost performance, decision-level fusion is also employed. When one of the sub-components is reporting that there is no data with which to make a decision, then the combined method defaults to the value of the remaining component.

Figure 5 demonstrates this combined method for person identification. When a speaker turns their head away from the camera, face recognition drops to 0, but the combined system defaults to speaker recognition and successfully identifies the target. When the target is too quiet for speaker recognition, then the combined method relies on face recognition. Finally, when the wrong person would have been selected by speaker recognition, score-level fusion with face recognition overcomes.

B. Multi-Speaker Interaction

For multi-speaker interactions, however, the described fusion algorithm is incomplete. For one, the camera field of view is too limited, and so the robot must be rotated to face the speakers in order for face recognition to contribute correctly to the combined person identification system. This step is accomplished with the speech localization system

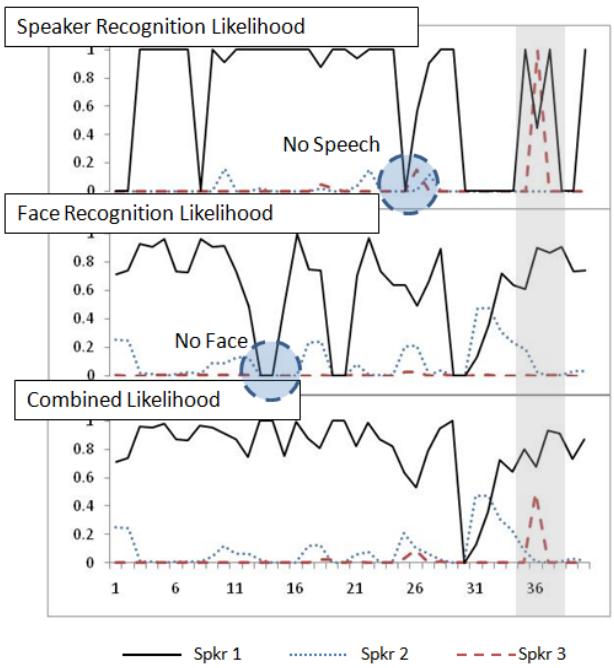

Fig 5. Likelihood results for each of the three systems (Speaker, Face, Combined), for a single recorded session. The combined likelihood improves performance over either speaker or face when one modality is missing. Furthermore, score level fusion corrects for a mistake in the speaker recognition (see grayed area).

described in Section IV.C. Our robot, Octavia, is capable of up to 50° of rotation to either side without rotating the Segway base.

Another challenge, however, which was demonstrated by the human-subjects experiment is the limitations on speaker set size for speaker recognition. A robot may talk to many more than 5 people in a day, but all of their verifiers cannot be run simultaneously with any reasonable precision. Even assuming multi-speaker conditions, the number of people talking to the robot at one time is much lower, and reasonably handled by speaker recognition. So the robot needs to load speakers into the set dynamically. Face recognition, in combination with sound source localization, provides an elegant means for accomplishing this task.

Face recognition is also limited in the number of faces it can be searching for simultaneously, but it can handle several times more potential candidates at one time then speaker recognition. Therefore, when face recognition strongly believes that a new speaker has entered the environment, a new speaker is dynamically included in the set of active speakers. This means that a new verifier is inserted with the new speaker model, and the imposter models for all existing speakers are updated to include the new speaker. Similarly, when a face has not been detected for a suitably long time (currently 5 minutes), that speaker can be removed from the set of active speakers and all imposter models updated accordingly.

The combined speaker tracking system with speech localization, robotic movement, and dynamic speaker set creation, is implemented as a finite state automata (see Figure 6). Speaker models are stored in a database for easy access and retrieval.

C. Discussion

The system can load speaker models from a database based on face recognition results. Can it also create new speaker models? The existing system can, in fact, create new speaker models in place of loading an existing model when a speaker does not yet exist in the database. The data collection challenges, however, are not insignificant. Currently, when a face is detected that the robot does not have a model for, the robot asks an open ended question such as, "How is your day going?" and records the audio stream. After either a maximum collection time, 20-sec, or no speech is detected for 5-sec, the robot creates a new model and stores it in the database.

Creating new models in this fashion, however, has its limitations. For one, it assumes that face recognition can detect the speaker, but there is no speaker model. This is not entirely unreasonable, as faces may have been learned somewhere other than on the robot itself, but that may be an uncommon state. More limiting, however, is that a speaker may not speak long enough, or loud enough to create a reasonable model, and without a true dialogue management system to get the user to speak for a longer period of time, the resulting speaker model will be very poor.

A solution to this problem, and to other problems relating to poor speaker model precision, is online learning, or updating, of the speaker model. If the fused system is reasonably certain of the current speakers identity, but speaker recognition for that speaker is generally poor (perhaps the speaker has a cold), then new data could be collected and integrated into the model. This does risk adding bad data to a model, which could weaken models even further. However, there is currently no reliable method for updating models, and there needs to be one. Fusing data across sensory mediums with different fail points, and limiting updates to individual verifiers, may mitigate this negative outcome.

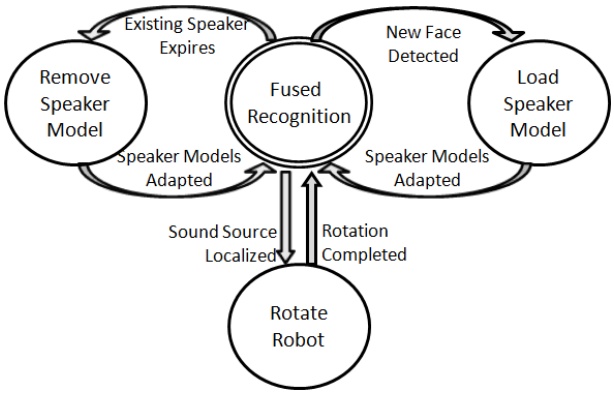

Fig 6. Finite state automata for following a multi-speaker dialogue.

VI. Conclusion

This work has presented results for person identification by a humanoid robot, including face recognition, speaker recognition, and a fused method combining both face and speaker recognition methods. Both face recognition and speaker recognition used established algorithms for identifying people, but, despite individually strong recognition rates, they can fail in a real-time, continuous interaction. This is because people are not always looking at the robot, a requirement for face recognition, and they are not always talking, a requirement for speaker recognition. By fusing the two classification systems together with both decision and score level fusion, we not only overcome missing results from either classifier, but correct for false positive results from individual classifiers.

VII. References

[1] A. D. Yarmey, A.L. Yarmey, M.J. Yarmey, and L. Parliament "Commonsense Beliefs and the Identification of Familiar Voices," Appl. Cognit. Psychol. 15,2001, pp. 283-299.

[2] V. Bruce, and A. Young. *In the Eye of the Beholder: The Science of Face Perception*. Oxford: Oxford University Press. 1998

[3] A. Schmidt-Nielsen, and T.H. Crystal. "Speaker Verification by Human Listeners: Experiments Comparing Human and Machine Performance Using the NIST 1998 Speaker Evaluation Data," *Digital Signal Processing* vol. 10 (1-3). January/April/July, 2000 pp. 249-266

[4] S. Campanella, and P. Belin, "Integrating Face and Voice in Person Perception", *Trends in Cognitive Sciences*, Volume 11, Issue 12, December 2007, pp. 535-543.

[5] S. Kiesler, "Fostering common ground in human-robot interaction". Proceedings of the IEEE International Workshop on Robots and Human Interactive Communication (RO-MAN 2005), Nashville, TN, August 2005, pp. 158-163.

[6] A. Powers, A.D. Kramer, S. Lim, J. Kuo, S.L. Lee, and S. Kiesler, "Eliciting information from people with a gendered humanoid robot." Proceedings of the IEEE International Workshop on Robots and Human Interactive Communication, Nashville, TN, August 2005, pp. 158-163.

[7] A. Powers and S. Kiesler, "The advisor robot: tracing people's mental model from a robot's physical attributes", Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction, March 02-03, 2006, Salt Lake City, Utah, USA

[8] E. Martinson and A. Schultz, "Discovery of sound sources by an autonomous mobile robot*" Autonomous Robots*, vol. 27 (3), pp. 221-237.

[9] M. Turk, A. Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience*, vol. 3(1), 1991.

[10] B. Draper, K. Baek, M. S. Bartlett, and J.R. Beveridge, "Recognizing faces with PCA and ICA", *Computer Vision and Image Understanding (CVIU),* July-August 2003, pp. 115-137.

[11] P Belhumeur, J Hespanha, and D Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection", European Conference on Computer Vision (ECCV), 1996.

[12] J. Wright, A.Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210-227, Feb. 2009.

[13] J. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, and H. Okuno, "Robust Recognition of Simultaneous Speech by a Mobile Robot," *IEEE Transactions on Robotics*, v. 23(4), pp742-752, 2007

[14] K. Nakadai, H. Nakajima, Y. Hasegawa, and H. Tsujino. "Sound source separation of moving speakers for robot audition," Int. Conf. on Acoustics, Speech and Signal Processing, 2009

[15] F. Krsmanovic, C. Spencer, D. Jurafsky, A. Ng, "Have we met? MDP Based Speaker ID for Robot Dialogue," 9th International Conf. on Spoken Language Processing, 2006

[16] T. Quatiri. *Discrete Time Speech Signal Processing*. Pearson Education Inc, Dehli, India. 2002

[17] M. Ji, S. Kim, H. Kim, "Text-Independent Speaker Identification using Soft Channel Selection in Home Robot Environments." IEEE Transactions on Consumer Electronics, v. 54(1), pp 140-144, 2008

[18] S. Palanivel, B. Yegnanarayana, "Multimodal Person Authentication using speech, face and visual speech*", Computer Vision and Image Understanding (CVIU)*, 2008, pp. 44--55.

[19] L. Lacheze Y. Guo, R. Benosman, B. Gas, and C. Couverture, "Audio/Video Fusion for Objects recognition", Intelligent Robots and Systems, St. Louis, MI, 2009

[20] B. Kamgar-Parsi, W. Lawson, B. Kamgar-Parsi, "Recognizing Faces like Humans", SPIE Newsroom, 22 February 2010.

[21] R. Cole, M. Noel and V. Noel, "The CSLU Speaker Recognition Corpus", Proc. of the Int Conf. on Spoken Language Processing, 1998