Deictic Pronoun Learning and Mirror Self-Identification

Kevin Gold
Yale University
51 Prospect St.
New Haven, CT, USA
kevin.gold@yale.edu

Brian Scassellati
Yale University
51 Prospect St.
New Haven, CT, USA
scaz@cs.yale.edu

Abstract

The ability to identify the self in a mirror reflection and the ability to use the word “I” effectively are commonly seen as major milestones in a human infant’s development of a concept of self. In addition, deictic pronouns such as “I” and “you” present a technical challenge to computational methods for grounded word learning, which have commonly associated word definitions with sensory patterns instead of pragmatic roles. Here, a robot learns the usage of the words “I” and “you” by observing others playing a game of catch, then correctly uses the terms to refer to its mirror image and to a conversational partner, respectively. Word learning occurs by using already understood words (“got the ball”) to infer the referents of spoken sentences. The properties of those referents, including the conversational roles of “speaker” and “addressee” as well as properties unique to each person, are then associated with the unknown words, and the significance of these associations ranked via chi-square tests. After sufficient observation of others using “I” and “you,” the robot’s own usage is correct without any need for supervised learning. To achieve mirror self-recognition, the robot uses the timing of the visual feedback that results from its arm’s movement. The part of the image that is labeled as “self” is then treated as the robot’s location in the image for the purpose of responding to the command, “Say who got the ball.”

1. Introduction

In human development, there are two major milestones associated with the ability to reason about the self. The first is the ability to correctly answer the question, “Who’s that in the mirror?” This ability typically does not develop until about two years of age (Amsterdam, 1972). Since answering this question requires the ability to use language, which may be distinct from self-recognition, a second measure called “the mirror test” was devised for chimpanzees (Gallup, 1970) and later adapted to human infants (Amsterdam, 1972). Typically, a mark that the subject cannot feel or see directly is applied to the subject while the subject is unconscious. Upon waking, its behavior in front of a mirror is observed. If the subject uses the mirror to direct its hands toward the mark, it is said to be “self-aware” by the standards of the mirror test. Since the test’s conception, the list of organisms that succeed in recognizing themselves in the mirror has grown to include chimpanzees, orangutans, and bonobos (Gallup et al., 2002); human infants beginning at 18 months (Amsterdam, 1972); and, in modified form, dolphins (Reiss and Marino, 2001).

Because of the rarity of animal species that can pass the mirror test, some researchers have concluded that a broad range of intelligent capabilities, particularly social understanding, are necessary to accomplish mirror self-recognition (Gallup et al., 2002). But this assumption has never been rigorously tested. It has also been unclear just how broad the divide is between being able to find a mark in a mirror, and being able to learn the word “I” and apply it to the mirror image.

Building a robot that can learn to identify its mirror image as “I” can address these questions, while also providing insight into how we should be building intelligent robots. Some of the difficult aspects of the mirror test are also difficult for modern robots. Identifying the robot’s own body parts when they are seen in an unexpected place is impossible for any robot with a preprogrammed kinematic model that tells it where to look for visual feedback. Learning that the word “I” refers to the speaker in general, and that it only refers to the robot when it is the speaker, is impossible for a robot that can only associate words with visual images. By reverse-engineering how humans learn to identify their mirror image as “I,” we can potentially get a glimpse of some useful principles of human intelligence.

This paper describes a robot that uses only a few preprogrammed assumptions to come to the conclusion that its mirror image should be called “I.” By combining our previous work on learning the meanings of the words “I” and “you” through observation (Gold and Scassellati, 2006b) with a module that learns to classify movement in the visual field as self-generated (Michel et al., 2004), the robot can achieve a task that it could not perform using either module separately:

the robot refers to its mirror image as “I,” without ever being explicitly trained to do so.

Indeed, the only training required is for the robot to observe the motion of its hand for a brief while, to learn when to expect visual self-feedback, and for the robot to observe two people tossing a ball back and forth saying “I got the ball” and “you got the ball” as appropriate. From these, the robot concludes that “I” refers to the speaker, “you” refers to the addressee, and that motion occurring roughly 500 ms after it sends a motor command is most likely itself. When it sees its ball in the mirror next to its reflection, the robot can then move, see that the reflection moved, identify the reflection as itself, conclude that the best word to refer to the reflection is “I,” and say, “I got the ball.”

2. Prior Work

The learning of “I” and “you” described here was originally inspired by a system presented in (Oshima-Takane et al., 1999), which learned in simulation that “I” referred to the speaker and “you” referred to the addressee. That neural network did not address the problem of how the system identified who a statement was about; the referent of the target word was given directly as an integer to the system. It also did not address the problem of word learning in the context of more than these two words, which made the interpretation of the results somewhat more difficult. Nevertheless, Oshima-Takane’s work showing that these words are not learned through one-on-one interaction, but by observing complete conversations (Oshima-Takane, 1992, Oshima-Takane et al., 1996) has been valuable in designing our “I” and “you” learning system (Gold and Scassellati, 2006b). The chi-square test has been used previously to find statistically significant collocations of words in text (Manning and Schütze, 1999), as well as to compute distance measures on image properties (Steels and Kaplan, 2002), but we know of no previous system to use chi-square tests to rank word-property associations in the manner presented here.

Our motion-based self-recognition module was originally presented in (Michel et al., 2004). The idea of using feedback time to find the self has been presented before (Fitzpatrick, 2003), but always in conjunction with some other kinematic model that would learn a specific expected location for the robot’s manipulator, then be unable to deal with the fact that the mirror image was not where it expected it to be.

Though other work has hard-coded the understanding of deictic pronouns such as “my” and “your” (Roy et al., 2004), that work did not present an approach to learning these words. The same lab has also performed a great deal of work on associating words with sensory properties (Roy and Pentland, 2002), but not conversational roles or similarly abstract properties.

This is our first paper to combine our self-recognition module and language-learning module to produce a robot that can produce utterances about itself, rather than merely interpreting the statements of others. It is also our first to add a representation of the robot itself to the word-learning environment, along with the ability to recognize when the robot is assuming the role of speaker or addressee.

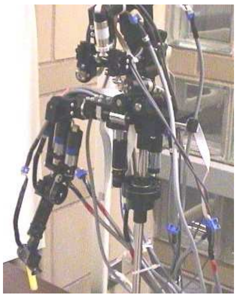

Figure 1: Nico, the physical robot that performed the self-recognition task.

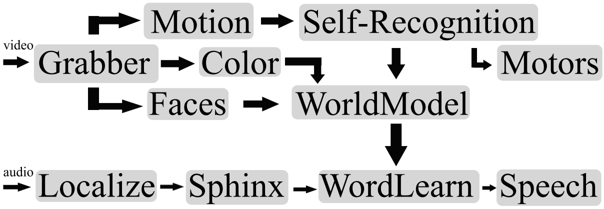

Figure 2: Processing pipelines for video (top) and audio (bottom). Modules that detect patches of bright color, faces, and self-generated motion work in parallel to provide the world model with labels for image regions. The self-recognition module also generates motor commands, using the timing of visual feedback to identify regions of the image under the robot’s own control. Audio is localized using the timing difference between the two microphone channels, then passed to speech recognition software. Once converted to text, the recognized speech is compared to the sensed environment to ground word meanings. If the recognized speech is found to be a command asking for information, the word learning module accesses word definitions and chooses a word that best matches the model for output.

3. Methods

3.1 *Robotic Architecture*

The robot Nico (Figure 1) consists of a 3 DOF eye assembly that controls a wide and narrow field-of-view CCD camera for each eye, a 4 DOF neck, and a six DOF arm and hand assembly, all connected via serial cables to a rack of 16 processors running the QNX Neutrino operating system. Typically one processing module runs on each node. The nodes are physically connected via Ethernet cable and a 100 Mbit switch, while in software different modules communicate via a socket-like communications system that allows either blocking or nonblocking reads by multiple consumer processes from the same producer process.

Visual information is sent from the cameras and processed at a rate of 30 frames per second. This information is then sent down three different pathways for processing. One pathway finds faces using the Intel OpenCV vision library, which in turn uses the boosted pattern-finding algorithm of (Viola and Jones, 2004) to accomplish face detection at a rate of about 1 FPS. A second pathway detects motion by finding absolute pixel differences from frame-to-frame, then performs smoothing and region-growing on these areas to create bounding boxes around areas of motion. A third pathway finds regions of high color saturation, which are smoothed and boxed in a manner similar to the motion module. (See Figure 2.)

The motion bounding boxes are filtered through a self-recognition module, which finds motion that coincides with the robot's motor commands. Then all three kinds of bounding boxes – faces, color, and motion – are passed to the world model, a simple two-dimensional representation of the world consisting of boxes in the visual field. Faces and self-labeled motion are treated as agents, while the color boxes are treated as objects that agents can "possess" by being closest to them.

This state of the world is sent over a wireless TCP/IP connection to a Windows laptop, where speech processing takes place using the Sphinx 4 speech recognition system (Walker et al., 2004). Audio is captured on the laptop using a powered dual-channel microphone; the individual left and right receivers were placed roughly one foot apart and three feet in front of the robot, so as to reduce motor noise. Crude localization, consisting of merely a "left" or "right" judgment, was performed by noting which channel registered the sound wave first. In this way, the statements interpreted via the Sphinx speech recognition system could be attributed to the agents in the visual field.

Speaker-independent speech recognition in Sphinx 4 is accomplished using a trained Hidden Markov model (HMM) that uses an established vocabulary to help drive phonological judgments. In this case, a simple context-free grammar (CFG) was used to make word decisions:

```
<utterance> = <subj> <verb> <obj>
<subj> = I | you | Alice | Bob | (say who)
<verb> = got | caught
<obj> = it | the ball
```

The state of the world, the incoming recognized language, and its speaker are then combined in the word-learning step, described below.

3.2 Word Learning

The word learning system relies on sentence context and previously understood words to learn the meanings of "I" and "you." On hearing a sentence, the system finds all already understood words in the sentence. It then searches the physical environment for agents that possess those properties. All unknown words in the sentence then become more strongly statistically associated with only the properties of those agents.

Properties of agents are modeled as Boolean variables, calculated deterministically by earlier modules in the pipeline. Property types can include actions ("speaking"), being the target of an action ("addressee," the target of a "speaking" action), unique identifiers, or other perceptual properties.

Grammatical parsing has not been implemented; thus, sentences are treated as simple collections of words. When a sentence is heard, each word for which an association has been learned with confidence adds its associated property to a list of properties to seek in the environment. The system then makes a list of agents satisfying any of these properties; for example, if it heard and understood "caught," it would look for agents that had recently caught something.

For each word-property pair, a chi-square value is calculated to determine the significance of the association. Typically, 2×2 chi-square tests assess the likelihood of the null hypothesis that two events are independent – in this case, the two events being whether a word is in a sentence, and whether the sentence refers to an agent that possesses the property. Here, the system adopts a method from statistical natural language processing, and simply ranks the properties for each word by their chi-square value to determine which property is best associated with the word. This method has previously been used to find words that appear together with high frequency (Manning and Schütze, 1999). Here, instead of word-word associations, the system calculates word-property associations.

Negative associations are excluded from analysis; that is, word-property pairs that occur significantly less often than expected are excluded, despite the fact that these pairs may achieve higher chi-square values overall than the positive associations. In addition, word-property associations are treated as unreliable until the expected value of each square in the chi-square table is at least 5; this is to avoid the problem of chi-square tests being unreliable when data is sparse (Manning and Schütze, 1999).

For its initial list of known words, the system allows certain words to be simply preprogrammed; it does not attempt to handle the case in which no words are yet learned. A full developmental system might incorporate pointing and gaze into that initial word-learning step. "I" and "you," however, are both learned by human children well after they have begun to learn other words (Messer, 1994), so it is reasonable to assume some already known words in our model.

The system is obviously naïve from a computational linguistics perspective, and is not meant to imply that grammar is unnecessary to word learning. For the experiments presented here, however, additional complexity was unnecessary given the simple context-free grammar from which words could be drawn.

In the experiments that follow, the following properties were used as possible meanings. *Speaker* was true of whoever was speaking, as determined by localization, while *Addressee* was true of the person being addressed. *HasBall* was true of whoever was closest to the ball, and the words "got" and "caught" were predefined as referring to this prop-

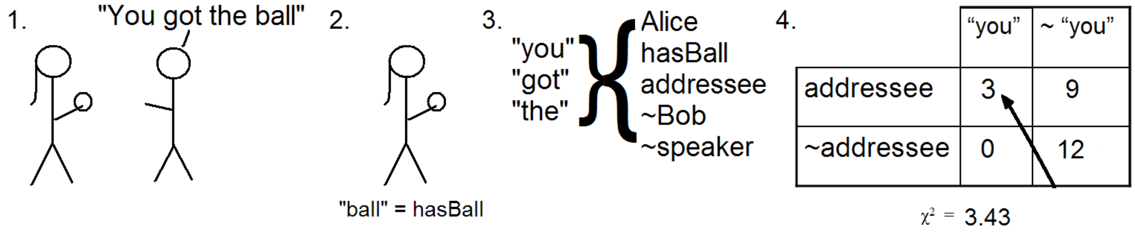

Figure 3: An illustration of the context-based word learning system, reprinted from (Gold and Scassellati, 2006a). (1) Alice says to Bob, "you got the ball." The speech recognition module turns this into a string. (2) Searching for words that are already understood, the system finds that "ball" refers to the property of *hasBall*, which is only true of Alice. (3-4) Alice's properties, but not Bob's, are used to update the chi-square tables representing the associations between each undefined word and each property.

erty. *LeftProperty*, *RightProperty*, and *NicoProperty* were assumed to be true of the person on the left side of Nico's field of vision, the person on the right side, and the robot itself, respectively. These abstract properties were proxies for any sensory attributes that were different for each participant and did not change over the course of the experiment, any of which would generate exactly the same chi-square values given the experiment setup.

For evidence in simulation that the system should scale to a larger number of dynamically changing properties and a real parent-child dialogue, see (Gold and Scassellati, 2006b).

3.3 *Self-identification*

Self-identification in the system uses the motion-based strategy presented in (Michel et al., 2004). In an exploratory phase prior to the experiment, the system would perform random arm motions within its field of view, and note the time elapsed between sending motor commands and receiving the resulting motion bounding boxes from the motion detection module. The round trip time between the decision to act and receiving the processed visual feedback follows a normal distribution with 95% of values falling between half a second and a second. This half-second window could then act as a filter on future motion bounding boxes, such that only objects that began to move within that time frame were classified as "self." After motion stops, the area enclosed by the self-labeled bounding box remains classified as "self" until the robot moves again.

3.4 *Speech production*

The only kind of productive utterance Nico currently makes is in response to the command, "Say who got the ball." In preparation for an utterance, Nico sets its own *Speaker* property to *true*, and sets the *Addressee* property of the person who gave the command to true. Nico then searches the utterance for already understood words, using the same function used to determine context described above. This produces the understood word "got" (or, equivalently, "ball"), associated with the property *HasBall*. Nico then searches the environment for an agent for which *HasBall* is true, and then finally searches its vocabulary for a word that does not mean *HasBall* but refers to a property that is true of the agent. If more than one word appears to apply, Nico uses the word with the highest chi-square value for its associated property. Thus, if the human who gave the command has the ball, and Nico does not know the speaker's name, Nico will find *Addressee* as the only property to which it can refer. Similarly, if the robot determines that it has the ball itself, it can use a word associated with *Speaker*. Nico then combines the found word with the phrase "got the ball" to answer the question.

If no person satisfies the *HasBall* property, Nico's preprogrammed behavior is to use "nobody." Similarly, if Nico has no word that can describe a property of the agent for whom *HasBall* is true, the preprogrammed response of "can't say" is used in place of an understood word.

4. Implementation and Results

The robot loaded chi-square values for word-property pairs it had learned from a previous experiment (Gold and Scassellati, 2006a), in which two people tossed a plastic yellow ball back and forth in front of Nico, commenting on who had the ball using utterances from the context-free grammar presented above. The data set used contained a total of 50 utterances. Using the property definitions generated by this data, the robot associated "you" with the property of *addressee* and "I" with the property of *speaker*.

Next, a mirror was placed in front of Nico, and Nico's self-recognition module was enabled. In addition, Nico was programmed to move its arm roughly every 20 seconds. The ball was either held by experimenter, who alternately stood either to the left or the right of the mirror, or placed on Nico's base, where it was only visible to Nico in its mirror reflection. (This setup is shown from Nico's point of view in Figure 4.) Alternating between these two locations for the ball, the experimenter told Nico, "Say who got the ball."

Under these conditions, Nico correctly answered "I got the ball" 16 out of 20 times (80%) when it could see the ball at the base of its mirror reflection. The incorrect answer "You got the ball" was given twice, and on two trials the command was simply misinterpreted as the state-

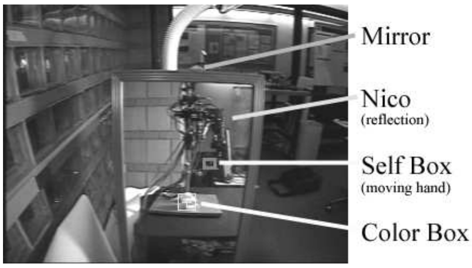

Figure 4: A view of the experimental setup, taken from one of Nico's cameras. The robot can see its reflection in the mirror, center. Superimposed on the image is the bounding box produced by the "color" module, which has found the mirror reflection of the bright yellow ball, and the box produced by the "self-motion" module, which has found the reflected motion caused by Nico's arm movement. Neither arm nor ball is currently visible to Nico except in the mirror. When the experimenter entered the field of view and commanded Nico to "say who got the ball," the robot used the word associations it learned while watching a game of catch to correctly reply "I got the ball" when the ball was placed as shown, and "you got the ball" when the experimenter held it.

ment "You got the ball." Moving once every twenty seconds proved to be confusing, however: the experimenter often happened to move within Nico's learned time window for self-recognition, and this motion was mistaken for self-motion. As a result, when the experimenter held the ball, the answer "You got the ball" was given only 11 out of 20 times (55%), with "I got the ball" given 7 times (35%) and incorrect speech recognition in two trials (10%).

When the behavior of the robot was changed so that it moved only once, rather than intermittently fidgeting throughout the experiment, the performance was much improved. "I got the ball" was given as a correct answer on 18/20 trials (90%), with one answer of "you got the ball" and one speech recognition error. "You got the ball" was the answer given on all 20 trials when the experimenter held the ball.

5. Conclusions

While Nico's world model and language capabilities are simplistic in several respects, this robot is the first to learn meanings for the words "I" and "you" from observation, then apply them successfully itself. In particular, its usage of the word "I" to refer to its own mirror image, when the meaning of that word had to be learned from observation, is an exciting step for robotics. At the same time, a considerable degree of caution and skepticism is necessary in interpreting this behavior. The term "self-awareness" is not terribly useful from a scientific point of view; the system does what it does. Instead, we might examine how the system achieves this behavior, and which elements were crucial to its success.

The first element critical to the system's performance was its ability to *identify actions and their targets*, and make them salient properties for word acquisition. Obviously, if the system were forced to associate words with specific visual patterns, deictic pronouns could not have been learned, because they can refer to different individuals. Less obviously, to learn "you" a property had to be added to the model that had nothing to do with the agent in isolation, but was true because of the way the agent was being acted on (spoken to) by another agent. The authors believe that associating words with functional roles and properties, rather than superficial appearance, will prove a profitable method for applying machine learning to semantics in the long term.

The second critical ability was that of being able to *learn language from observed conversation*, rather than from robot-directed speech. Without the chance to observe shifting deictic referents, the system would not have had sufficient examples to infer the meaning of "you," which would have always referred to the robot – though it might have learned "I" given interactions with different speakers.

Critical to this ability to learn language from observation was its ability to *infer referents from context*, using known words to narrow down the possible referents for the unknown words. During every utterance, there was always a speaker and an addressee, so the mere presence of both properties in the environment would have been insufficient to distinguish "I" and "you." By using context to narrow down the possible referents, the system was able to determine which of the two people was being talked about, and did not encounter the troubles that would have befell a naïve "everything with everything" associationist system.

Then, to use the words effectively, the robot had to *equate its own representation and properties with those of human agents*. It is not necessarily obvious to a robotic system that when a robot speaks, it is performing the same action as when a human speaks; in fact, this equivalence had to be explicitly programmed. Otherwise, the robot would know that "I" can refer to human speakers, but would never use the word itself, because it would not know that in speaking, it was assuming an analogous role to a human speaker. The same equivalence was also implicit in representing the robot's self-feedback as an agent in the environment. In Nico, these equivalences were preprogrammed, but a more general ability to map human actions and properties onto the robot's equivalent states would be a desirable goal for research.

Finally, to achieve mirror self-recognition, the robot had to *recognize novel visual feedback as self-generated*. For many practical applications, this ability is totally unnecessary. A preprogrammed kinematic and physical model that works only when the robot's arm where it is expected to be will often work just fine to achieve the robot's goals. In fact, this may well be why the ability to recognize oneself in the mirror is so sparsely distributed in the animal kingdom: why waste neurons on the ability to recognize novel self-generated feedback, when a system with built-in parameters can work reliably with less training? Gallup's mirror test (Gallup, 1970) may not be a test of self-awareness so much as it is a test of flexibility in recognizing feedback. Yet this ability certainly confers benefits to humans: it allows us to drive cars, use computer mice, and yes, comb our hair in front of mirrors.

Nico possesses only the barest of functionality from each of the areas listed above, and each component offers much room for improvement. Action recognition, language learning, inferring meaning from context, and adaptive control are each active areas of research, and in any one of these areas more complex systems than Nico already exist. Yet, by combining elements of each of these technologies, and with a few key insights about the nature of language and self-recognition, Nico has exhibited a behavior that has elsewhere been considered a major milestone in human development. While the "mirror test" as performed on nonhuman species perhaps does not imply as much about other aspects of intelligence as previously thought – most of the abilities cited above were related to the robot's learned usage of "I," and not mere mirror self-recognition – it would appear that asking a child "who's that in the mirror?" truly does test several different aspects of intelligence that are important to how we define intelligence in humans.

Acknowledgments

Support for this work was provided by a National Science Foundation CAREER award (#0238334). Some parts of the architecture used in this work was constructed under NSF grants #0205542 (ITR: A Framework for Rapid Development of Reliable Robotics Software) and #0209122 (ITR: Dance, a Programming Language for the Control of Humanoid Robots) and from the DARPA CALO/SRI project.

References

Amsterdam, B. (1972). Mirror self-image reactions before age two. *Developmental Psychobiology*, 5(4).

Fitzpatrick, P. (2003). *From first contact to close encounters: a developmentally deep perceptual system for a humanoid robot*. PhD thesis, MIT.

Gallup, G., Anderson, J., and Shillito, D. (2002). The mirror test. In Bekoff, M., Allen, C., and Burghardt, G., (Eds.), *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*. MIT Press.

Gallup, G. G. (1970). Chimpanzees: self-recognition. *Science*, 167(3914):86–87.

Gold, K. and Scassellati, B. (2006a). Grounded pronoun learning and pronoun reversal. In *Proceedings of the Fifth International Conference on Development and Learning*, Bloomington, IN.

Gold, K. and Scassellati, B. (2006b). Using context and sensory data to learn first and second person pronouns. In *Human-Robot Interaction 2006*, Salt Lake City, Utah.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Messer, D. J. (1994). *The Development of Communication*. Wiley, West Sussex, England.

Michel, P., Gold, K., and Scassellati, B. (2004). Motion-based self-recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sendai, Japan.

Oshima-Takane, Y. (1992). Analysis of pronominal errors: a case study. *Journal of Child Language*, 19.

Oshima-Takane, Y., Goodz, E., and Derevensky, J. L. (1996). Birth order effects on early language development: do secondborn children learn from overheard speech? *Child Development*, 67:621–634.

Oshima-Takane, Y., Takane, Y., and Shultz, T. (1999). The learning of first and second person pronouns in English: network models and analysis. *Journal of Child Language*, 26:545–575.

Reiss, D. and Marino, L. (2001). Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence. *Proceedings of the National Academy of Sciences*, 98(10).

Roy, D., Hsiao, K.-Y., and Mavridis, N. (2004). Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, 34(3).

Roy, D. K. and Pentland, A. P. (2002). Learning words from sights and sounds: a computational model. *Cognitive Science*, 26:113–146.

Steels, L. and Kaplan, F. (2002). Aibo's first words: The social learning of language and meaning. *Evolution of Communication*, 4(1).

Viola, P. and Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., and Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition. Technical Report TR-2004-139, Sun Microsystems.