ARL-TR-8898 ● JAN 2020

A Review of Artificial Intelligence (AI) Algorithms for Sound Classification: Implications for Human–Robot Interaction (HRI)

by Troy Kelley and Kelly Dickerson

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-8898 ● JAN 2020

A Review of Artificial Intelligence (AI) Algorithms for Sound Classification: Implications for Human–Robot Interaction (HRI)

Troy Kelley
Human Research and Engineering Directorate, CCDC Army Research Laboratory

Kelly Dickerson
CCDC Data & Analysis Center

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
January 2020	Technical Report	October 2018–September 2019

4. TITLE AND SUBTITLE
A Review of Artificial Intelligence (AI) Algorithms for Sound Classification: Implications for Human–Robot Interaction (HRI)

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Troy Kelley and Kelly Dickerson

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
CCDC Army Research Laboratory
ATTN: FCDD-RLH-FD
2800 Powder Mill Road, Adelphi, MD 20783-1138

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-8898

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This report presents a review of artificial intelligence (AI) algorithms and their application to audition in a human–robot interaction (HRI) context. The AI algorithms selected for auditory perception ultimately have an impact on computational transparency, system behavior explainability, and ultimately, the quality of the HRI. AI algorithms applied to auditory perception include sounds sensed and processed by a software system, as well as sounds emitted by a software system that are meant to be recognized by a human listener. Some major classes of AI algorithms, specifically neural networks, deep learning, hidden Markov models, and hybrid models will be reviewed in the context of machines' sound processing. Additionally, the effects of each class of algorithm on transparency and HRI will be discussed. Recent work in AI algorithm development suggests that hybrid models may be the best approach for sound processing as they are recommended for complex data processing and decision-making. Hybrid models blend approaches to maximize the benefits while minimizing the limitations of multiple techniques. A set of general recommendations are included in the final section of the report.

15. SUBJECT TERMS
artificial intelligence, human–robot interaction, transparency, hybrid architectures, neural networks

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Susan Hill
Unclassified	Unclassified	Unclassified	UU	28	19b. TELEPHONE NUMBER (Include area code) (410) 278-6237

Contents

1. Introduction 4

1.1 Definitions of AIs, Algorithms, Agents, and other Thinking Machines 4

1.2 Benefits of Sound 4

1.2.1 Using Speech to Interact With Technology 5

1.2.2 Beyond Speech: Environmental Sound 5

1.2.3 Sound Stimulus Preparation 7

2. Major AI Approaches 7

2.1 Current State of the Art of AI 8

2.2 Neural Networks 9

2.3 Deep Learning 11

2.4 Markov Models 12

2.5 Hybrid Models 13

3. Discussion 14

4. Conclusion and Future Directions 16

5. References 17

List of Symbols, Abbreviations, and Acronyms 25

Distribution List 26

1. Introduction

1.1 Definitions of AIs, Algorithms, Agents, and other Thinking Machines

This report describes the advantages and disadvantages of some of the major approaches to artificial intelligence (AI), including neural networks, deep learning, hidden Markov models (HMM), and hybrid approaches in developing algorithms for the classification of auditory information. Before discussing how sound can enhance situational awareness (SA), it is important to define some of the common terms surrounding the discussion of AI. An AI algorithm is a collection of software functions that allow the AI to discriminate and label various sensory inputs such as sounds or images. AI algorithms can also make decisions based on those inputs. The combination of AI and physical sensing capabilities is an embodied intelligent machine. Examples of embodied intelligent machines include many different consumer and military products—devices such as Amazon Echo, Tesla vehicles, or a robotic vehicle-gunnery station. In the military context, AI algorithms are often implemented in a mobile robotics context (i.e., intelligent agents) and embedded in mobile systems, such as the high-mobility multipurpose wheeled vehicle, and even embedded in a stationary Tactical Operations Center (TOC). Intelligent agents can be considered a subcategory of embodied intelligent machines. While there are multiple salient feature differences between possible types of embodied intelligent machines, the critical distinction for the purpose of this report is between AI algorithms (software) and embodied intelligent machines (software + sensing and action). Finally, an autonomous system refers to robotic assets (agents) and other embodied intelligent machines that can act entirely independently from their human teammates.

1.2 Benefits of Sound

Soldier tasks tend to rely predominantly on the visual modality unless the task is explicitly communication related. Vehicle operators, pilots, unmanned aerial vehicle operators, and data workers, such as command and control TOCs and cybersecurity experts, all have visually demanding roles and are at risk of visual overload. Moving information from one modality to another is one strategy for mitigation of visual workload. For example, pilots with high visual workload can use auditory cueing to improve performance and SA (Calhoun et al. 1987). In addition to reducing visual load by serving as another information channel, auditory information can enhance SA (Endsley 1995); for example, a sudden increase or decrease in the loudness of traffic could suggest the presence of a threat, or a change

in the frequency spectrum could indicate the presence of more than just passenger vehicles on a roadway. Evidence from the multisensory integration literature suggests that in day-to-day life, humans will dynamically shift attention between auditory and visual information, maximally weighing information from either sounds or objects (depending on the context), their prior experience, and perceived reliability (or unreliability) of those cues (Shams and Seitz 2008). Despite the clear benefit of auditory information to Soldier SA, Army autonomous systems have generally underutilized sound as a potential source of information about the environment. Current autonomous systems do not listen, but rather rely on text-based communication from Soldiers to gain an understanding about mission goals, environment status, and team status (ARL 2017).

1.2.1 Using Speech to Interact With Technology

There are several reasons why the Army has relied primarily on text inputs rather than moving toward speech for robotic systems and intelligent agents. There is familiarity and a history of using text-based systems; and automatic speech recognition (ASR) systems, while a significant technical innovation, have several potential usability and signal processing issues when considered in an operational context. ASR technologies are not new (see Haridas et al. 2018); however, recent iterations using AIs (i.e., Google Duplex; see Leviathan and Matias 2018) have improved the ability of the system to recognize both the phonetic and semantic contents of speech. Despite the capabilities improvements of AI-supported ASRs, these technologies are not really well suited to the battlefield. Universally, old and new ASRs require a low signal-to-noise ratio to function properly. Pilots, vehicle operators, or dismounted Soldiers would be hard pressed to find an area quiet enough that they could effectively use voice commands to interact with their vehicles or other technologies. Further, speech can be easily overheard, making the situations where voice commands could be used limited to cockpits and other vehicle crew areas.

1.2.2 Beyond Speech: Environmental Sound

ASR systems have limited practical utility in an operational context because of noise and the need to keep sensitive communications quiet. Environmental sound, non-speech auditory stimuli produced by the activities of animals, humans, and machines, can provide critical information for squads to leverage to improve SA. Familiar sounds and the soundscapes they create in urban environments create a representation of "normal" or "safe" operations. Changes to ambient-sound level or the objects contributing to the auditory milieu can convey important information. When detected, these changes can improve overall SA. However, human listeners often miss changes in the auditory environment (change deafness; see Gregg and

Samuel 2008). Sensors can be used to "hear" sounds earlier than a human could detect the event and AIs may be able to assist in recognition of important changes in the ambient auditory environment by classifying sounds and then alerting Soldiers to mission-relevant information. Using AI to detect and classify sounds reduces the impact of known human-perceptual limitations such as change deafness, inattentional deafness (Gregg and Samuel 2008), and informational masking (Dickerson and Gaston 2014), and represents the potential for a synergistic relationship between the sensing and classification abilities of the machine with the decision-making capabilities of the human. The remainder of this report is devoted to highlighting some of the recent work in environmental sound classification and providing background and perspective of common AI frameworks.

1.2.2.1 Environmental Sound Classification

Classification is the algorithmic process used to predict category membership for items in a data set. Classification is also a foundational capability for any AI. Thus, we will use classification performance as a running example through Section 3. Automatic classification of auditory scenes is still relatively novel compared to visual object classification. In a review of the auditory scene classification literature, Salamon and Bello (2015) found few examples of scene classification, and many of those were restricted to using global scene properties and generally did not attempt to extract and classify individual sound objects within a scene. This limitation is not unique to auditory scene classification. A similar challenge exists for visual objects embedded in scenes. Naturalistic scenes are noisy compared to single images and sounds, and while a scene provides beneficial context for a human, for machine classifiers it can cause problems. Humans are able to segment individual auditory streams (Mondor 1994) and visual objects (Helmholtz 1925; Hochberg 1981) from a complex background and this is a fundamental aspect of object recognition. However, this human-like object classification process, regardless of the modality, has been a significant challenge for AI. Adding to the special challenge of sound recognition, urban scenes can unfold in almost any acoustic configuration and, unlike speech and music, urban auditory scenes do not have any sort of higher-order rule structure (i.e., grammar, composition) that could serve as an initial organizing principle. Urban planners and designers map and classify the sound objects contributing to a scene; however, this process has relied overwhelmingly on "low-tech" methods, such as manual sound file annotation and interviews with city residents. These inefficient processes may be acceptable for humans trying to understand a particular sound environment; however, for automatic classification and AI training, "low-tech" data sets means small data sets with limited a priori tagging. These two factors make creation of model training

sets particularly difficult for sound (see Vogiatzis and Remy 2018 for review of methods for mapping urban soundscapes).

1.2.3 Sound Stimulus Preparation

In terms of stimulus processing for AI applications, sound classification research proceeds much like visual object recognition. Sounds are segmented from the background, preprocessed in various ways, and filtered to approximate the human sensory and perceptual processes. The most common filter choice in recent years for sound is the mel-frequency cepstrum (MFC), likely because it approximates the psychological sensation of frequency height of a pure tone (see Serizel et al. 2017). Earlier efforts tended to rely on rectangular or critical bands (Green et al. 1984); however, these filters were poorly suited to complex stimuli such as environmental sounds.

Irrespective of the sensory stream, after filtering, transformed features are extracted and these features are then fed into an AI mechanism for classification. The specific features extracted and submitted for classification vary depending on the particular approach selected (see Salamon and Bello 2015 and Serizel et al. 2017 for detailed reviews). Once classification is accomplished, retraining of the network can be an issue.

2. Major AI Approaches

AI has made great strides since its early days. The AI community has been researching problem solving since the 1960s (Minsky 1961). Initial efforts on problem solving were largely in the domain of chess using symbolic systems (Simon and Simon 1962). The domain of chess as a problem space turned out to be relatively well defined. Even in the 1950s, it was understood that chess was a relatively easy problem space. As Shannon (1950) noted: The chess machine is an ideal one to start with since

1) the problem is sharply defined in terms of the allowed operations (i.e., the moves) and in the desired goal (i.e., checkmate);

2) it is neither so simple as to be trivial nor too difficult for satisfactory solution;

3) chess is generally considered to require "thinking" for skillful play; a solution of this problem will force us to either to admit the possibility of mechanized thinking or to further restrict our concept of "thinking"; and

4) the discrete structure of chess fits well into the digital nature of modern computers.

All of these points from Shannon allowed for the application of symbolic approaches. In other words, the symbolic approaches used in the era of Shannon's chess machine were inherently discrete and precisely quantifiable. For example, language and mathematics are symbolic systems and early implementations of these systems within AI were referred to as "production systems". These systems were similar to formal predicate logic systems using "if–then" formalisms. However, it was previously discovered by Kurt Godel in 1931 that symbolic systems are inherently incomplete (Hofstadter 1979) and thus are not capable of accounting for every eventuality in a logical space. To complicate AI research, and what fostered the early overenthusiasm, was that these early symbolic systems did not account for some of the more difficult aspects of stimulus perception since chess pieces are stationary, identifiable, and consistent. As AI progressed, there was a move away from chess toward visual perception as a problem space, and an attempt to begin to approximate the variety and complexity of the real world in the problem spaces and training sets used in AI applications (Horn 1986).

2.1 Current State of the Art of AI

Current AI systems have achieved superhuman levels of performance in various games including Go (Lee 2019a; Lee 2019b), chess (Campbell et al. 2002), and backgammon (Tesauro 1994). Current AI approaches, like the early focus on chess, also tend to focus on specific problem domains as opposed to developing a more generalized learning system (Kelley and Long 2010). One notable example of a system that made the jump from domain specific to potentially generalizable is Watson (the supercomputer program) created by IBM to play Jeopardy. Watson only played Jeopardy and did not play other types of games, such as Wheel of Fortune. As the Watson platform matured, the system became a more sophisticated question-and-answer expert system performing a task more like a search engine and less like a Jeopardy contestant. Watson and similar AI systems start out solving one type of problem, but have been applied outside their original domains; for example, Watson has been found especially useful for medical diagnostics, which can be a constrained problem space and expert systems lend themselves well to the task. Chen et al. (2016) found that Watson's ability to organize billions of pages of textual information and create novel textual connections was extremely useful for certain domains, like drug repurposing and novel drug candidate selection.

Many AI solutions to classification problems rely upon the "brute force" approach to problem solving, which relies on computer power to examine large numbers of

possibilities in a vast search space. Indeed, the brute-force methodology implemented in Watson and many currently used deep learning architectures rely on sheer processing power to work with the huge sets of training stimuli to develop a reliable AI. In recent years, advancements in graphical processing units (GPUs) allowed deep learning architectures to take advantage of additional computational power and thus, GPU improvements in the accuracy of machine classifiers using brute-force approaches to object classification. In 2011, researchers used GPUs and deep learning to achieve superhuman performance in a visual pattern recognition task involving handwritten digits (Ciresan et al. 2011). While speech recognition performance begins to approximate human performance, sound classification still lags behind, with the best-performing classifiers scoring between 40%–70% depending on the specifics of the approach (see Salamon and Bello [2015] for example classification accuracies and Moffat et al. [2017] for review).

In the next sections, four major AI methodologies are discussed in relation to the classification problem. Specifically, neural networks (Section 2.2), deep learning (Section 2.3), HMM (Section 2.4), and hybrid approaches (Section 2.5) are discussed. While there are benefits and limits to each approach, the present review discusses these factors in the context of autonomous systems for improving SA and representational methodologies for improved computational transparency. For example, deep learning AI architectures produce a learned output; however, the representational network supporting a particular learning outcome is not transparent and may be fundamentally different from the representational network constructed by a human mind when presented with the same information. For autonomous systems that operate independently, this is not an issue; however, for teaming, a shared representation is critical for building shared SA and, ultimately, beneficial functional autonomous systems that can operate effectively in the dynamically changing conditions of a battlefield.

2.2 Neural Networks

Neural networks (McClelland and Rumelhart 1986) are distributed classification systems that are meant to act much like the distributed collections of neurons in the brain, except that the information in a neural network is largely static, with discrete calculations at each node, while the human synapses are dynamic, changing constantly as when new experiences are added. Neural networks have been used extensively for the classification of a variety of stimuli including images (Krizhevsky et al. 2012), handwritten characters (LeCun et al. 1990), and more recently, sound (Piczak 2015).

Neural networks operate by extracting important information (feature vectors) to classify data. In auditory stimuli, pitch, loudness, and sound duration are just some of the features that a neural network researcher would attempt to extract as feature vectors to classify the data. Auditory researchers frequently use the MFC, which is a power-spectrum representation of sound and a convenient feature to analyze. However, a problem with developing neural networks for sound is that, unlike images where features are often spatially separated and have face validity (a line is recognized as a line but also part of a larger object), sounds almost always arrive to the listener comingled, where a given auditory stream contains parts of multiple sound sources.

Feature extraction for automated systems must proceed in a manner similar to humans where the arriving cacophony is segregated based on both low-level feature similarity (i.e., frequency, intensity, direction) and then grouped to form meaningful sound representations based on cognitive factors such as prior experience and listening context.

Vision researchers have traditionally used image databases, which are static representations of the world (Deng et al. 2009), because they are easy to work with and low in data requirements. However, static data ignore the additional information from dynamic changes in the world. Newer research shows that temporal information in video improves classification and reduces the number of training examples required (Simonyan and Zisserman 2014) and using deep learning techniques for action recognition in videos further increases improvements (Sharma et al. 2015). Auditory researchers have found that static, neatly parsed, data can differ greatly from the more dynamic, real-world spoken data. For example, people often speak in short bursts and phrases, and emit nonlinguistic sounds (ah, hum) that are not well interpreted by current speech-to-text systems (Anagnostopoulos et al. 2015). This makes development of real-world sound classification systems difficult.

As part of the AI development cycle, neural network researchers sometimes use "black-box" techniques to extract features from a data set (Castelvecchi 2016). Black-box techniques create trained systems where the decision-making process is not immediately apparent. For example, a decision-tree system allows the decision-making process to be broken down into a series of steps and decision points, allowing the consumer of the information to check each decision at each point in the network.

However, black-box techniques do not create a decision tree. For instance, within a neural network, the information related to the decision-making process is stored as a collection of weights across the nodes in the network, making explanations of

the decision-making process nearly impossible. So black-box techniques yield classification systems that are not explainable by a human and can reduce the efficiency of the subsequent Human–Robot Interaction (HRI). In other words, the features extracted by black-box techniques are not necessarily known to the neural network developer, and the use of those features for classification yields unexplainable results, even to the developer of the system.

This black-box problem where feature extraction results may not be knowable, even by developers, led to recent research in explainable AI, which should be of benefit to the HRI community. For example, when a decision is made by an AI system, the consumer of that information might need additional information about how the decision was made and want to know the AI's confidence in the decision. Further, for HRI in particular, a decision made by the AI must be transparent and understandable (see Gunning 2017 and Gunning and Aha 2019 for discussion on explainable AI). These types of concerns may seem basic, but the decisional processes and transparency of those system decisions are critical in high-risk domains such as medical diagnoses, aviation support, and battlefield SA. Explainable AI is especially helpful in the after-action reports and debriefings that accompany military operations. Systems like Debrief, which was a hybrid system, were used to justify actions for the TacAir-Soar combat domain (Laird et al. 1994) and explainable AI systems aided in military planning and execution (Tate et al. 2000).

2.3 Deep Learning

Deep learning is a neural network technique, or set of techniques, that is currently popular in the AI community. Deep learning has generally shown better performance as compared to neural networks for some types of classification tasks (LeCun et al. 2015). For example, in 2011 researchers used deep learning techniques to achieve superhuman performance in a visual pattern-recognition task involving handwritten digits (Ciresan et al. 2011). However, the better performance for deep learning compared to traditional neural network approaches could be due to the use of brute-force training methodologies. The brute-force approach can lead the network to overgeneralize. Overgeneralization (i.e., overfitting) is the phenomenon where the model is highly tuned for accuracy on the training set, but performs poorly on novel examples at test (Srivastava et al. 2014). Overfitting has been a known problem with neural networks for a long time (Tetko et al. 1995). This also leads to another problem where a novel input cannot be rejected from a trained set of data. In other words, the neural network does not know that it does not know an instance of given data.

Other neural network problems persist as well. For example, deep learning neural nets have classified nonsense images as trained images (Nguyen et al. 2015). And while deep learning has been applied to auditory data for speech recognition with typical improvements over traditional neural networks (Amodei et al. 2016), it has recently been shown that deep learning for speech recognition can be fooled by nonsense data in the same ways as image classification (Cisse et al. 2017). This could have enormous implications for secure data neural networks in the Department of Defense. If neural networks trained to recognize specific auditory commands could be easily fooled by nonsense vocalizations, or vocalizations that appear to be correct, neural networks could miss obvious environmental sound classifications on the battlefield or take actions that are not intended by the designers.

In terms of human interaction, deep learning has some of the same problems as traditional neural networks. If feature extraction is done by a deep learning network, then the lack of explainability from a black-box process can be even more detrimental to humans interacting with the AI. Furthermore, issues of overtraining and overgeneralization can make the learning achieved by deep learning processes nontransferable to other situations, perhaps without the understanding by the end user of the system—meaning that the system was intended to be used on specific data only and is not valid outside of that constrained data set.

2.4 Markov Models

The Markov decision-making process, and the family of Markov decision-making models, has been used in robotics for several decades (Koenig and Simmons 1998) and is extremely useful for speech recognition (Haridas et al. 2018) and human motion trajectory prediction (Rudenko et al. 2019) by modeling the state emission probabilities (Renals et al. 1994). In terms of AI classification systems, speech recognition is probably one of the best applications of HMMs.

Speech recognition algorithms benefit from having a time series probability distribution, something that is not applicable in static visual domain sets. For example, in speech recognition, each word cues the probability of the next word in a sentence, so a set of probabilities for each word in a sentence is easily defined in a computationally efficient manner. This problem space has lent itself well to the use of HMMs. HMMs were used extensively in the 1980s and caused a rapid advancement in the field of speech recognition (Haridas et al. 2018). For speech recognition tasks, the state of the world can be defined in finite terms, but for other more difficult tasks, like environmental sound recognition, the state space is much more vast, or nondeterministic polynomial (NP) time, known as NP-complete. By

NP-complete we mean the time required to solve such problems is so vast that it is essentially prohibitive to the efficient solving of the problem.

Unlike deep learning neural networks, which rely on extracting features for classifications, HMMs instead predict state transitions, which can be explained to the end user and thus improve the overall human interaction with the AI. For instance, it can be explained that a certain word in a sequence of words was the most likely next word in a sentence, and that is why the word was predicted as an outcome. This is in contrast to neural networks, where even the designer of the network might not understand why a network arrived at a certain decision. Further, HMMs have been used to estimate the reward functions of human operators, and consequently select behaviors that are congruent with the perceived policy of the human operators (Tabrez and Hayes 2019). This application of HMM makes any communication with the human operators more seamless since understanding of reward functions is relatively straight forward and easily understood by human users.

Recently, HMM architectures have been combined with deep learning in a hybrid methodology, which is discussed in the next section. This hybrid union of techniques allows the time-dependent aspects of HMM to be combined with the classification strengths of deep learning (Hinton et al. 2012).

2.5 Hybrid Models

As technology becomes more complex, and as AI solutions are used in more complex decision-making areas, there is an increased need to integrate and unify architectures with an emphasis on transparency and explainability to the end user. In general, no single type of AI algorithm offers the best solution for all possible classification problems, including auditory classification problems. Instead, the best solution for difficult computational problems is usually to combine algorithm classes and theoretical approaches into a more seamless hybrid approach. A hybrid approach leverages the strengths of each computational methodology into a unified whole, while avoiding the pitfalls associated with each individual approach.

As neural networks proved less effective than HMMs for speech processing, there was a shift in AI approaches in the late 2000s to develop more hybrid systems of both neural networks and HMMs (Juang and Rabiner 2004). With respect to specific problems in sound, hybrid systems have been used to advance the state of the art in speech recognition (Hinton et al. 2012). These hybrid systems use a mixture of deep learning and HMM. Additionally, hybrid models were used to advance the state of the art for recognition of emotion in speech (Pao et al. 2007).

This theoretical point of using hybrid methodologies has been argued for several decades in the AI community, especially for combining neural network approaches with more traditional symbolic approaches (Smolensky 1987; Holyoak 1991; Sun and Alexandre 1997; Sun 2001; Kelley 2003; Anderson et al. 2004; Clark and Pulman 2007; Jilk et al. 2008). A hybrid approach has also been used to successfully develop several cognitive architectures capable of complex problem solving including Soar (Laird et al. 1994), the Atomic Components of Thought-Rational (ACT-R) (Anderson et al. 2004), and Clarion (Sun 2006). These architectures are goal-based symbolic systems at the high level, and human-like memory decay systems at the low level (Anderson 2005).

In a recent review of hybrid models, Fajardo-Toro et al. (2018) argue that as the complexity of problems increases, the need for models to handle more complex, dynamic, nonlinear relationships also increases, and this need necessitates the continued development of hybrid models. For example, newer hybrid models in speech emotion recognition apply deep learning and HMMs (Li et al. 2013). Dao et al. (2019) recently argued that hybrid models, in the areas of structural engineering, were needed because earlier single-method models could not answer more complex problems, specifically, determining the properties of geo-polymer concrete. Recently, hybrid models have shown promise for solving more complex problems than models based on single AI algorithms across a variety of different problem domains. A survey of hybrid models in finance showed that hybrid methodologies outperformed other methods of classification in a variety of complex financial problem spaces including bankruptcy prediction and financial forecasting (Bahrammirzaee 2010). Additionally, hybrid models have been used to better estimate possible human trajectories (Rudenko et al. 2019) for oil price prediction (Wu et al. 2019) and wildfire prediction (Jaafari et al. 2019)

3. Discussion

In summary, discrete symbolic approaches originally used to play chess in the 1960s were replaced by distributed neural networks in the 1980s, which emerged as a new approach in AI. Since then, deep learning neural networks have used a variation on multilayer processing, or convolution, to add computational power to traditional neural networks. Additionally, increased computational speed for neural network training is provided by recent advancements in the design of GPUs. Multiple GPUs are inexpensive for researchers to use on a small scale and can be operated in parallel for increased computational power.

However, the current deep learning methodology is a brute-force AI approach; it uses large sets of training data and multilayer networks for learning and does not

develop transparent learning methods that can be easily communicated to human users. Additionally, deep learning neural networks can be fooled by nonsense data (Radford et al. 2015) and there are issues of overfitting and lack of generalization (Spigler 2019). Other architectures, specifically HMMs, have been used with success in speech-to-text systems. HMM researchers benefited from the temporal nature of speech, which allows the efficient use of HMM systems in speech recognition. While HMMs perform relatively well for speech, and while speech recognition automation has made great strides, other types of sound classification problems still persist. The best HMM, deep learning, and neural network approaches all put up classification accuracy values in the 40%–80% range, depending on the size of the data set, the length of the samples, and the particulars of the classification approach.

Hybrid approaches, like the Symbolic and Sub-symbolic Robotics Intelligence System (Kelley 2006) or ACT-R/Leabra (O'Reilly and Munakata 2000) both combine neural networks with production systems for solving complex classification and decision-making problems. For classification problems, hybrid systems can use deep learning neural networks to process large amounts of data at the lower levels, combined with higher levels of predicate logic for decision making, and the integration of ontologies for contextualized problem solving. This use of the higher-level information and context could potentially improve classification accuracies for environmental sound sets, which are often smaller and more abstract than the image sets used previously in deep learning, neural network, or HMM approaches. Jaafari et al. (2019) warns that the benefits of hybrid models over traditional single-model approaches might not be evident during initial training, but that hybrid models were shown to be more robust during the validation phases of development and less susceptible to overfitting and overtraining. This is something for researchers to keep in mind during any model comparison.

In terms of computational transparency and explainability, symbolic systems are, by nature, more intuitively explainable than distributed systems—as language itself is an example of a symbolic system. The use of ontologies ConceptNet (Liu and Singh 2004) and Cyc (Lefkowitz et al. 2007) with spoken interactions from the user would allow increased understanding given the symbolic nature of the representation. However, while current hybrid models offer a solution for communication and HRI using speech, there is currently no research examining environmental sound perception for AI systems using these architectures or a comparable hybrid approach that leverages the neural networks' ability to handle huge amounts of data while using higher-level cognitive concepts to refine a solution and support better classification and decision making, explainability, and ultimately, SA. This is an important gap in the literature and mirrors a gap that

existed until the late 1990s in the human environmental sound perception literature. Research on acoustic and semantic influences on sound perception was entirely separate. Gregg and Samuel (2008) found that both acoustic and semantic information influenced sound detection. More recently, researchers are using ontologies to incorporate context into natural language dialog to improve overall conceptual understanding (Rajpathak et al. 2012). However, few models of environmental classification, as opposed to language processing, are relying on transformations of acoustic data and do not incorporate any higher level contextual information.

4. Conclusion and Future Directions

Additional research on environmental sound processing, perception, and cognitive decision making using sound is needed for humans, agents, and human–agent teams to develop the best sensing, algorithmic, and communication strategies for autonomous systems to increase SA in complex urban environments.

Use of hybrid AI methodologies (e.g., neural networks with HMM) may provide the best performance, and work to improve transparency and create learning that is semantically similar to human learning is needed to improve SA at a team or squad level.

The HRI community needs to be aware and understand the limitations of AI techniques, especially in the areas of overtraining/overgeneralization and black-box techniques. These techniques can be detrimental to increased transparency, which is needed to ensure adequate HRI.

Humans learn to speak and understand speech by extracting regularities from the auditory environment. Environmental sound perception proceeds much the same way. Future research developing auditory classification training sets should focus on developing a strategy for understanding how noise in the data can benefit learning; noisy data in the right context can create robust and highly transferable learning.

5. References

Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, Casper J, Catanzaro B, Cheng Q, Chen G, et al. Deep speech 2: end-to-end speech recognition in English and Mandarin. In: Balcan MF, Weinberger KQ, editors. ICML 2016. Proceedings of the 33rd International Conference on Machine Learning, vol. 48; 2016 June 19–24; New York City, NY. J Mach Learn Res; c2016. p. 173–182.

Anagnostopoulos CN, Iliou T, Giannoukos I. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif Intell Rev. 2015;43(2):155–177.

Anderson JR. Cognitive psychology and its implications. 6th ed. New York (NY): Worth Publishers; 2005.

Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, Qin Y. An integrated theory of the mind. Psychol Rev. 2004;111(4):1036–1060.

[ARL] Army Research Laboratory (US). Robotics Collaborative Technology Alliance (RCTA) proposed 2017–18 biennial program plan. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2017.

Bahrammirzaee A. A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. Neural Comput Appl. 2010;19(8):1165–1195.

Calhoun GL, Valencia G, Furness TA III. Three-dimensional auditory cue simulation for crew station design/evaluation. In: Proceedings of the Human Factors Society 31st Annual Meeting; 1987 Sep; New York, NY. Santa Monica (CA): Human Factors and Ergonomics Society. 1987;31(12):1398–1402.

Campbell M, Hoane AJ, Hsu F. Deep blue. Artif Int. 2002;134(1–2):57–83.

Castelvecchi D. Can we open the black box of AI? Nature News. 2016;538(7623):20.

Chen Y, Argentinis E, Weber G. IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. Clin Ther. 2016;38(4):688–701.

Ciresan DC, Meier U, Gambardella LM, Schmidhuber J. Convolutional neural network committees for handwritten character classification. ICDAR 2011. 2011 International Conference on Document Analysis and Recognition; 2011

Sep 18–21; Beijing, China. Los Alamitos (CA): IEEE Computer Society; c2011. p. 1135–1139.

Cisse M, Adi Y, Neverova N, Keshet J. Houdini: fooling deep structured prediction models. 2017 July. arXiv preprint arXiv:1707.05373.

Clark S, Pulman S. Combining symbolic and distributional models of meaning. In: Quantum Interaction, Papers from the 2007 AAAI Spring Symposium, Technical Report SS-07-08; 2007 Mar 26–28; Stanford, CA. Menlo Park (CA): AAAI Press; c2007. p. 52–55.

Dao DV, Trinh SH, Ly H-B, Pham BT. Prediction of compressive strength of geopolymer concrete using entirely steel slag aggregates: novel hybrid artificial intelligence approaches. Appl Sci. 2019;9(6):1113.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei, L. ImageNet: a large-scale hierarchical image database. In: CVPR 2009. 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 June 20–25; Miami Beach, FL. Piscataway (NJ): Institute of Electrical and Electronic Engineers, Inc. p. 248–255.

Dickerson K, Gaston JR. Did you hear that? The role of stimulus similarity and uncertainty in auditory change deafness. Front Psychol. 2014;5:1125.

Endsley MR. Toward a theory of situation awareness in dynamic systems. Hum Fact. 1995;37(1):32–64.

Fajardo-Toro CH, Mula J, Poler R. Adaptive and hybrid forecasting models—a review. In: Engineering digital transformation—lecture notes in management and industrial engineering. Cham (Switzerland): Springer; 2018. p. 315–322.

Green DM, Mason CR, Kidd G Jr. Profile analysis: critical bands and duration. J Acoust Soc Am. 1984:75(4):1163–1167.

Gregg MK, Samuel AG. Change deafness and the organizational properties of sounds. J Exp Psychol Hum Percept Perform. 2008;34(4):974–991.

Gunning D. Explainable artificial intelligence. 2017 Nov 17 [accessed 2020 Jan 14]. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

Gunning D, Aha DW. DARPA's Explainable artificial intelligence (XAI) program. AI Mag. 2019;40(2):44–58. https://doi.org/10.1609/aimag.v40i2.2850.

Haridas AV, Marimuthu R, Sivakumar VG. A critical review and analysis on techniques of speech recognition: the road ahead. Int J Knowl Int Eng Sys. 2018;22(1):39–57.

Helmholtz H. Physiological optics, vol III: the perceptions of vision. Southall JPC, translator and editor. Rochester (NY): Optical Society of America; 1925.

Hinton G, Deng L, Yu D, Dahl G, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Kingsbury B, Sainath T. Deep neural networks for acoustic modeling in speech recognition. IEEE Sig Proc Mag; 2012;29:82–97.

Hochberg J. On cognition in perception: perceptual coupling and unconscious inference. Cognition. 1981;10(1–2):127–134.

Hofstadter DR. Gödel, Escher, Bach: an eternal golden braid. New York (NY): Vintage Books; 1979.

Holyoak KJ. Symbolic connectionism: toward third-generation theories of expertise. In: Ericsson KA, Smith J, editors. Toward a general theory of expertise: prospects and limits. New York (NY): Cambridge University Press; c1991. Ch. 12, p. 301–329.

Horn BKP. Robot vision. Cambridge (MA): MIT press; 1986.

Jaafari A, Zenner EK, Panahi M, Shahabi H. Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability. Agr Forest Meteorol. 2019;266:198–207.

Jilk DJ, Lebiere C, O'Reilly RC, Anderson JR. SAL: an explicitly pluralistic cognitive architecture. J Exp Theor Artif Intell. 2008;20(3):197–218.

Juang BH, Rabiner LR. Automatic speech recognition–a brief history of the technology development. Atlanta (GA): Georgia Institute of Technology; 2004.

Kelley TD. Developing a psychologically inspired cognitive architecture for robotic control: the symbolic and sub-symbolic robotic intelligence control system (SS-RICS). Int J Advan Robot Sys. 2006;3(3):32.

Kelley TD. Symbolic and sub-symbolic representations in computational models of human cognition: what can be learned from biology? Theor Psychol. 2003;13(6);847–860.

Kelley TD, Long LN. Deep Blue cannot play checkers: the need for generalized intelligence for mobile robots. J Robot. 2010.

Koenig S, Simmons RG. Xavier: a robot navigation architecture based on partially observable Markov decision process models. In: Kortenkamp D, Bonasso RP, Murphy R, editors. Artificial intelligence based mobile robotics: case studies of successful robot systems. Cambridge (MA): MIT Press; 1998. p. 91–122.

Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. NIPS 2012. Proceedings of Advances in Neural Information Processing Systems 25; 2012 Dec 3–8; Lake Tahoe, Nevada. San Diego (CA): Neural Information Processing Systems Foundation, Inc.; c2012. p. 1090–1098.

Laird JE, Jones RM, Nielsen PE. Coordinated behavior of computer generated forces in TacAir-Soar. In: Collected papers of the Soar/IFOR project. Pittsburgh (PA): Carnegie Mellon University; 1994 Apr 25. Report No.: AD-A280 063. p. 57–64.

LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436.

LeCun Y, Boser BE, Denker JS, Henderson D, Howard RE, Hubbard WE, Jackel LD. Handwritten digit recognition with a back-propagation network. In: Touretzky DS, editor. Proceedings of Advances in Neural Information Processing Systems. San Diego (CA): Neural Information Processing Systems Foundation, Inc. 1990;2:396–404.

Lee C-S, Wang M-H, Chen L-C, Nojima Y, Huang T-X, Woo J, Kubota N, Sato-Shimokawara E, Yamaguchi T. A GFML-based robot agent for human and machine cooperative learning on game of Go. In: CEC 2019. Proceedings of the 2019 IEEE Congress on Evolutionary Computation. 2019 June 10–13; Wellington, New Zealand. Piscataway (NJ): IEEE Computer Society; 2019a. p. 793–799.

Lee C-S, Wang M-H, Ko L-W, Tsai B-Y, Yang S-C, Lin L-A, Lee Y-H, Ohashi H, Kubota N, Shuo N. PFML-based semantic BCI agent for game of Go learning and prediction. 2019b. arXiv:1901.02999.

Lefkowitz L, Curtis J, Witbrock M. Accessible research Cyc. Rome (NY): Air Force Research Laboratory (US); 2007 Sep. Report No.: AFRL-IF-RS-TR-2007-204.

Leviathan Y, Matias Y. Google Duplex: an AI system for accomplishing real-world tasks over the phone. Google AI Blog; 2018 May 8.

Li L, Zhao Y, Jiang D, Zhang Y, Wang F, Gonzalez I, Valentin E, Sahli H. Hybrid deep neural network—hidden Markov model (DNN-HMM) based speech

emotion recognition. In: 2013 CII. Proceedings of 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. 2013 Sep 2–5; Geneva, Switzerland. Los Alamitos (CA): IEEE Computer Society; c2013. p. 312–317.

Liu H, Singh P. ConceptNet—a practical commonsense reasoning tool-kit. BT Technol J. 2004;22(4):211–226.

McClelland JL, Rumelhart DE, PDP Research Group. On learning the past tense of English verbs. In: Parallel distributed processing, vol. 2. Explorations in the microstructure of cognition: psychological and biological models. Cambridge (MA): MIT Press; 1986. p. 216–271.

Minsky M. Steps toward artificial intelligence. Proceedings of the IRE. 1961;49(1):8–30.

Moffat D, Ronan D, Reiss J. Unsupervised taxonomy of sound effects. In: DAFx-17. Proceedings of the 20th International Conference on Digital Audio Effects. 2017 Sep 5–9; Edinburgh, United Kingdom.

Mondor TA, Bregman AS. Allocating attention to frequency regions. Percept Psychol. 1994;56(3):268–276.

Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: CVPR 2015. Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition; 2015 June 7–12; Boston, MA. Washington (DC): IEEE Computer Society; c2015. p. 427–436.

O'Reilly RC, Munakata Y. Computational explorations in cognitive neuroscience: understanding the mind by simulating the brain. MIT press: Cambridge (MA); 2000.

Pao TL, Chien CS, Chen YT, Yeh JH, Cheng YM, Liao WY. Combination of multiple classifiers for improving emotion recognition in Mandarin speech. In: Liao B-Y, Pan J-S, Jain LC, Liao M, Noda H, Ho ATS, editors. Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007); 2007 Nov 26–28; Kaohsiung, Taiwan. Los Alamitos (CA): IEEE Computer Society; c2007. Vol. 1, p. 35–38.

Piczak KJ. Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP); 2015 Sep 17–20; Boston, MA. Piscataway (NJ): IEEE Signal Processing Society; c2015. p. 1–6.

Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. 2015 Nov. arXiv preprint arXiv:1511.06434.

Rajpathak D, Chougule R, Bandyopadhyay P. A domain-specific decision support system for knowledge discovery using association and text mining. Knowl Inf Syst. 2012;31(3):405–432.

Renals S, Morgan N, Bourlard H, Cohen M, Franco H. Connectionist probability estimators in HMM speech recognition. IEEE T Speech Audi P. 1994;2(1):161–174.

Rudenko A, Palmieri L, Herman M, Kitani KM, Gavrila DM, Arras KO. Human motion trajectory prediction: a survey. 2019 May. arXiv preprint arXiv:1905.06113.

Salamon J, Bello JP. Unsupervised feature learning for urban sound classification. In: Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Brisbane, Australia; 2015 Apr 19–24. Piscataway (NJ): IEEE Signal Processing Society; c2015. p. 171–175.

Serizel R, Bisot V, Essid S, Richard G. Acoustic features for environmental sound analysis. In: Virtanen T, Plumbley MD, Ellis D, editors. Chapter 4, Computational Analysis of Sound Scenes and Events. Cham, Switzerland: Springer Verlag; c2018. p. 71–101.

Shams L, Seitz AR. Benefits of multisensory learning. Trends Cogn Sci. 2008;12(11):411–417.

Shannon CE. XXII. Programming a computer for playing chess. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1950;41(314):256–275.

Sharma S, Kiros R, Salakhutdinov R. Action recognition using visual attention. 2015 Dec. arXiv preprint arXiv:1511.04119.

Simon HA, Simon PA. Trial and error search in solving difficult problems: evidence from the game of chess. Behav Sci. 1962;7(4):425–429.

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014 Sep. arXiv preprint arXiv:1409.1556.

Smolensky P. Connectionist AI, symbolic AI, and the brain. Art Intel Rev. 1987;1(2):95–109.

Spigler G. Denoising autoencoders for overgeneralization in neural networks. IEEE T Pattern Anal. 2019 May 21.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–1958.

Sun R, Alexandre F, editors. Connectionist-symbolic integration: from unified to hybrid approaches. Hove, East Sussex (United Kingdom): Psychology Press; 1997.

Sun R. Artificial intelligence: connectionist and symbolic approaches. In: Smelser NJ, Baltes PB, editors. International encyclopedia of the social & behavioral sciences. Amsterdam (Netherlands): Elsevier Ltd; c2001. p. 783–789.

Sun R. The CLARION cognitive architecture: extending cognitive modeling to social simulation. In: Sun R, editor. Cognition and multi-agent interaction: from cognitive modeling to social simulation. New York (NY): Cambridge University Press; 2006. p. 79–100.

Tabrez A, Hayes B. Improving human–robot interaction through explainable reinforcement learning. In: HRI '19. Proceedings of the 2019 14th ACM/IEEE International Conference on Human–Robot Interaction (HRI); 2019 Mar 11–14; Daegu, Korea. Piscataway (NJ): IEEE Robotics & Automation Society; c2019. p. 751–753.

Tate A, Levine J, Jarvis P, Dalton J. Using AI planning technology for Army small unit operations. In: Chien S, Kambhampati S, Knoblock CA, editors. AIPS 2000. Proceedings of the Fifth International Conference on Artificial Intelligence Planning Systems; 2000 Apr 14–17; Breckenridge, CO. Palo Alto (CA): AAAI Press; c2000. p. 379–386.

Tesauro G. TD-Gammon, a self-teaching backgammon program, achieves master-level play. Neural Comput. 1994;6(2):215–219.

Tetko IV, Livingstone DJ, Luik AI. Neural network studies, 1. Comparison of overfitting and overtraining. J Chem Inf Comput Sci. 1995;35(5):826–833.

Vogiatzis K, Rémy N. Changing the urban sound environment in Greece: a guide based on selected case studies of strategic noise maps (SNM) and noise action plans (NAP) in medium and large urban areas. Environments. 2018;5(64). doi:10.3390/environments5060064.

Wu J, Chen Y, Zhou T, Li T. An adaptive hybrid learning paradigm integrating CEEMD, ARIMA and SBL for crude oil price forecasting. Energies. 2019;12(7):1239.

List of Symbols, Abbreviations, and Acronyms

ACT-R	Atomic Components of Thought-Rational
AI	artificial intelligence
ARL	Army Research Laboratory
ASR	automatic speech recognition
CCDC	US Army Combat Capabilities Development Command
GPU	graphical processing unit
HMM	hidden Markov models
HRI	Human–Robot Interaction
MFC	mel-frequency cepstrum
NP	nondeterministic polynomial
SA	situational awareness
TOC	Tactical Operations Center

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA

1 (PDF) CCDC ARL FCDD RLD CL TECH LIB

1 (PDF) CCDC ARL FCDD RLH B T DAVIS BLDG 5400 RM C242 REDSTONE ARSENAL AL 35898-7290

1 (PDF) CCDC ARL FCDD HSI J THOMAS 6662 GUNNER CIRCLE ABERDEEN PROVING GROUND MD 21005-5201

1 (PDF) USAF 711 HPW 711 HPW/RH K GEISS 2698 G ST BLDG 190 WRIGHT PATTERSON AFB OH 45433-7604

1 (PDF) USN ONR ONR CODE 341 J TANGNEY 875 N RANDOLPH STREET BLDG 87 ARLINGTON VA 22203-1986

1 (PDF) USA NSRDEC RDNS D D TAMILIO 10 GENERAL GREENE AVE NATICK MA 01760-2642

1 (PDF) OSD OUSD ATL HPT&B B PETRO 4800 MARK CENTER DRIVE SUITE 17E08 ALEXANDRIA VA 22350

ABERDEEN PROVING GROUND

11 (PDF) CCDC ARL FCDD RLH J LANE Y CHEN P FRANASZCZUK K MCDOWELL K OIE FCDD RLH BD D HEADLEY FCDD RLH FA A DECOSTANZA FCDD RLH FB A EVANS FCDD RLH FC J GASTON FCDD RLH FD A MARATHE S HILL