Artificial Intelligence Review
https://doi.org/10.1007/s10462-022-10224-2

A survey on deep reinforcement learning for audio-based applications

Siddique Latif[1] · **Heriberto Cuayáhuitl**[2] · **Farrukh Pervez**[3] · **Fahad Shamshad**[4] · **Hafiz Shehbaz Ali**[5] · **Erik Cambria**[6]

Abstract
Deep reinforcement learning (DRL) is poised to revolutionise the field of artificial intelligence (AI) by endowing autonomous systems with high levels of understanding of the real world. Currently, deep learning (DL) is enabling DRL to effectively solve various intractable problems in various fields including computer vision, natural language processing, healthcare, robotics, to name a few. Most importantly, DRL algorithms are also being employed in audio signal processing to learn directly from speech, music and other sound signals in order to create audio-based autonomous systems that have many promising applications in the real world. In this article, we conduct a comprehensive survey on the progress of DRL in the audio domain by bringing together research studies across different but related areas in speech and music. We begin with an introduction to the general field of DL and reinforcement learning (RL), then progress to the main DRL methods and their applications in the audio domain. We conclude by presenting important challenges faced by audio-based DRL agents and by highlighting open areas for future research and investigation. The findings of this paper will guide researchers interested in DRL for the audio domain.

Keywords Deep learning · Reinforcement learning · Speech recognition · Emotion recognition · (Embodied) dialogue

✉ Siddique Latif
siddique.latif@usq.edu.au

[1] University of Southern Queensland, Springfield, Australia

[2] University of Lincoln, Lincoln, UK

[3] National University of Science and Technology, Islamabad, Pakistan

[4] Information Technology University, Lahore, Pakistan

[5] EmulationAI, Brisbane, Australia

[6] Nanyang Technological University, Singapore, Singapore

Published online: 02 July 2022

1 Introduction

Artificial intelligence (AI) has gained widespread attention in different areas of research including computer vision, natural language processing (NLP), robotics, healthcare, and especially in audio signal processing. Audio processing covers many diverse fields including speech, music and environmental sound processing. In all these areas, AI techniques are playing crucial roles in designing audio-based intelligent systems (Purwins et al. 2019). One of the prime goals of AI is to create fully autonomous audio-based intelligent agents that can listen or interact with their environments to improve their behaviour over time through trial and error. Designing such autonomous systems has been a long-standing problem, ranging from robots that can react to the changes in their environment, to purely software-based agents that can interact with humans using natural language and multimedia. Reinforcement learning (RL) (Sutton et al. 1998) represents a principled mathematical framework of such experience-driven learning. Although RL had some successes in the past (Kohl and Stone 2004; Ng et al. 2006; Singh et al. 2002), previous methods were inherently limited to low-dimensional problems due to lack of scalability. Moreover, RL also has issues of memory, computational and sample complexity—in the case of learning algorithms (Strehl et al. 2006). Recently, deep learning (DL) models have risen as new tools with powerful function approximation and representation learning properties to solve these issues.

The advent of DL has dramatically improved the state-of-the-art performance and significantly impacted many areas from transportation to health and from social science to biology. Deep models such as deep neural networks (DNNs) (Hinton et al. 2012; Mohamed et al. 2009), convolutional neural networks (CNNs) (LeCun et al. 1989), and long short-term memory (LSTM) networks (Hochreiter and Schmidhuber 1997) have also enabled many practical applications by outperforming traditional methods in audio signal processing. The use of DL algorithms within RL has accelerated the progress of RL. This has given rise to the field of deep reinforcement learning (DRL). DRL embraces the advancements in DL to establish the learning processes, performance and speed of RL algorithms. This enables RL to operate in high-dimensional state and action spaces to solve previously unsolvable complex problems. Inspired by previous works such as (Lange et al. 2012), two outstanding works kick-started the revolution in DRL. The first was the development of an algorithm that could learn to play Atari 2600 video games directly from image pixels at a superhuman level (Mnih et al. 2015). The second success was the design of the hybrid DRL system, AlphaGo, which defeated a human world champion in the game of Go (Silver et al. 2016). In addition to playing games, DRL has also been explored in a wide range of applications such as computer vision (Le et al. 2021), natural language processing (NLP) (Naeem et al. 2020), robotics to control policies (Levine et al. 2016); generalisable agents in complex environments with meta-learning (Duan et al. 2016; Wang et al. 2016); indoor navigation (Zhu et al. 2017), and many more (Arulkumaran et al. 2017). In particular, DRL is also gaining increased interest in audio signal processing.

In audio processing, DRL has been recently used as an emerging tool to address various problems and challenges in automatic speech recognition (ASR), spoken dialogue systems (SDSs), speech emotion recognition (SER), audio enhancement, music generation, and audio-driven controlled robotics. In this work, we, therefore, focus on covering the advancements in audio processing by DRL. In Fig. 1, we present the cumulative distribution of publications in core DRL and applied to the audio domain. We note an emerging increased interest in the communities of both core and applied DRL. While core DRL grew

from 3 to 4 orders of magnitude from 2015 to 2021, applied DRL grew from 2 to 3 orders of magnitude in the same period.

There are multiple survey articles on DRL. For instance, Arulkumaran et al. (2017) presented a brief survey on DRL by covering seminal and recent developments in DRL—including innovative ways in which DNNs can be used to develop autonomous agents. Similarly, in Li (2017), authors attempted to provide comprehensive details on DRL and cover its applications in various areas to highlight advances and challenges. Other relevant works include applications of DRL in communications and networking (Luong et al. 2019), human-level agents (Nguyen et al. 2017), and autonomous driving (Sallab et al. 2017). None of these articles has focused on DRL applications in audio processing as highlighted in Table 1. This paper aims to fill this gap by presenting an up-to-date literature review on DRL studies in the audio domain, discussing challenges that hinder the progress of DRL in audio, and pointing out future research areas. We hope this paper will help researchers and scientists interested in DRL for audio-driven applications.

This paper is organised as follows. A concise background of DL and RL is provided in Sect. 2, followed by an overview of recent DRL algorithms in Sect. 3. With those foundations, Sect. 4 covers recent DRL works in the domains of speech, music, and environmental sound processing; and their challenges are discussed in Sect. 5. Section 6 summaries this review and highlights future pointers for audio-based DRL research and Sect. 7 concludes the paper.

2 Background

2.1 Deep learning (DL)

Deep neural networks (DNNs) have been shown to produce state-of-the-art results in audio and speech processing due to their ability to distil compact and robust representations from large amounts of data. The first major milestone was significantly increasing the accuracy of large-scale automatic speech recognition (ASR) using fully connected DNNs and deep autoencoders around 2010 (Hinton et al. 2012). It focuses to use DNNs with multiple non-linear modules arranged hierarchically in layers to automatically discover suitable representations or features from raw data. These non-linearities allow DNNs to learn complicated manifolds in speech and audio datasets. Various other deep architectures have shown the potentials to learn from audio to perform different tasks. Below we discuss these DL architectures, which are illustrated in Fig. 2.

Convolutional neural networks (CNNs) are a kind of feedforward neural networks that was specifically designed for processing images (Krizhevsky et al. 2012). However, CNNs have also been applied to various other fields including NLP (Arora et al. 2019), audio processing (Latif et al. 2019), and text analysis (Wang et al. 2019), and they have shown state-of-the-art performance. CNNs consist of a series of convolutional layers interleaved with pooling layers, followed by one or more dense layers. Whilst pooling layers reduce the spatial size of feature maps to decrease the amount of parameters, the neurons in dense layers are connected to every neuron in the preceding layer. In contrast to DNNs, CNNs limit the number of parameters and memory requirements dramatically by leveraging on two key concepts: *local receptive fields* and *shared weights*. A local receptive field refers to the region of the layer that is connected to any particular neuron in the next layer—note that receptive field, kernel and filter are used interchangeably. Shared weights, on the other

Table 1 Comparison of our paper with that of the existing DRL-based surveys

References	Focus			Details
	DRL	Audio Applications	Other Applications	
Arulkumaran et al. (2017)	✓	✗	✗	This paper presents a brief overview of recent developments in DRL algorithms and highlights the benefits of DRL and several current areas of research
Li (2017)	✓	✗	✓	This paper presents a generalised overview of recent exciting achievements of DRL and discusses core elements and mechanisms. It also discusses various fields where DRL can be applied
Luong et al. (2019)	✓	✗	communications and networking	This paper presents a comprehensive literature review on the applications of DRL in communications and networking, highlights challenges, and discusses open issues and future directions
Kiran et al. (2020)	✓	✗	autonomous driving	This paper summarises DRL algorithms and autonomous driving, where (D)RL methods have been employed.It also highlights key challenges towards real-world deployment of autonomous cars
Haydari and Yilmaz (2020)	✓	✗	transportation systems	This paper summarises existing works in the field of transportation, and discusses the challenges and open questions regarding DRL in transportation systems
Ours (2022)	✓	✓	✗	We present a comprehensive review focused on DRL applications in the audio domain, highlight existing challenges that hinder the progress of DRL in audio, and discuss pointers for future research

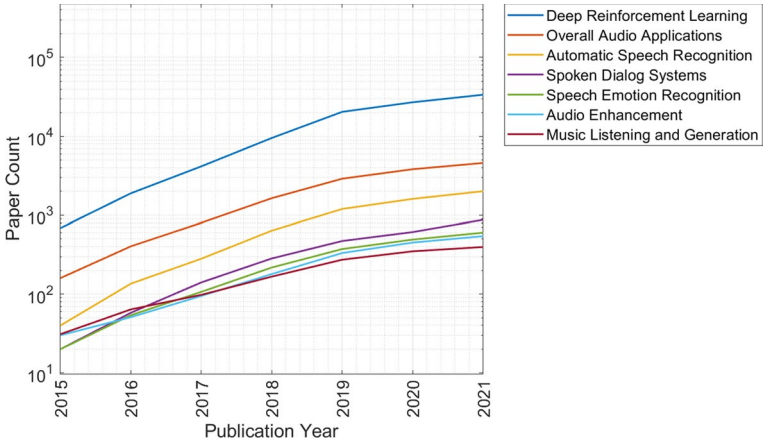

Fig. 1 Cumulative distribution of publications per year (data gathered from 2015 to 2021)—from https://www.scopus.com

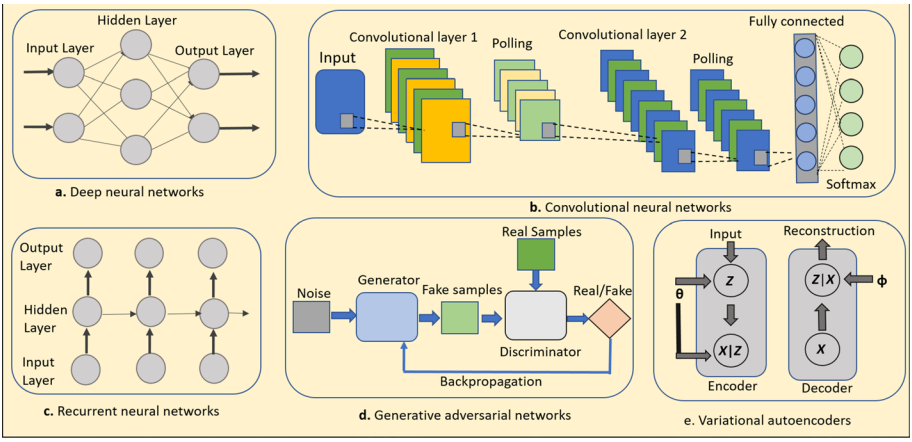

Fig. 2 Graphical illustration of different DL architectures

hand, refers to the same weights used across all receptive fields in same layer of CNN, as opposed to each receptive field in the layer having its own set of weights. Recently, CNN-based models have been extensively studied for a variety of audio processing tasks including music onset detection (Schlüter and Böck 2014), speech enhancement (Mamun et al. 2019), ASR (Abdel-Hamid et al. 2014), speech emotion recognition (Latif et al. 2020), etc. However, a raw audio waveform with high sample rates might have problems with limited receptive fields of CNNs, which can result in deteriorated performance. To handle this performance issue, dilated convolution layers can be used in order to extend the receptive field by inserting zeros between their filter coefficients (Chang et al. 2018; Chen et al. 2019).

Recurrent neural networks (RNNs) follow a different approach for modelling sequential data (Lipton 2015). They introduce recurrent connections to enable parameters to be shared across time, which makes them very powerful in learning temporal structures from the input sequences (e.g., audio, video). They have demonstrated their superiority over

traditional HMM-based systems in a variety of speech and audio processing tasks (Latif et al. 2020). Due to these abilities RNNs architectures including long-short term memory (LSTM) (Hochreiter and Schmidhuber 1997) and gated recurrent unit (GRU) (Cho et al. 2014) networks have an enormous impact in the speech community and have been incorporated in state-of-the-art audio-based systems. Recently, RNNs have also extended to include information in the frequency domain besides temporal information in the form of frequency-LSTMs (Li et al. 2015) and time-frequency LSTMs (Sainath and Li 2016). In order to benefit from both neural architectures, CNNs and RNNs can be combined into a single network with convolutional layers followed by recurrent layers, often referred to as convolutional recurrent neural networks (CRNN). Related works have shown CRNNs abilities in ASR (Qian et al. 2016), SER (Latif 2020), music classification (Ghosal and Kolekar 2018), and other audio related applications (Latif et al. 2020).

Sequence-to-sequence (Seq2Seq) models were motivated due to problems requiring sequences of unknown lengths (Sutskever et al. 2014). Although they were initially applied to machine translation, they can be applied to many different applications involving sequence modelling. In a Seq2Seq model, while one RNN reads the inputs in order to generate a vector representation (the *encoder*), another RNN inherits those learnt features to generate the outputs (the *decoder*). Seq2Seq models have been gaining much popularity in the speech community due to their capability of transducing input to output sequences. DL frameworks are particularly suitable for this direct translation task due to their large model capacity and their capability to train in an end-to-end manner—to directly map the input signals to the target sequences (Zhang et al. 2017; Lu et al. 2016; Liu et al. 2019). Various Seq2Seq models have been explored in the speech, audio and language processing literature including recurrent neural network transducer (RNNT) (Graves 2012), monotonic alignments (Raffel et al. 2017), listen, attend and spell (LAS) (Chan et al. 2016), neural transducer (Jaitly et al. 2016), recurrent neural aligner (RNA) (Raffel et al. 2017), and transformer networks (Pham et al. 2019), among others. In particular, transformer-based models have achieved unprecedented success in numerous speech and audio processing tasks including audio classification (Chi et al. 2021), speaker recognition (Wang et al. 2021), and speech-to-text (Bae et al. 2021), to name a few. These transformer models consist of an encoder-decoder architecture and work by leveraging the multi-head self-attention mechanism to consider the longer-distanced context around a word in a computationally efficient way (Vaswani et al. 2017). This makes them not only pay equal attention to all the elements in the sequence to boost accuracy but also results in harnessing the power of modern GPUs parallel environment for faster sequence processing compared to RNNs (Karita et al. 2019). For a more in-depth discussion about applications of transformers in audio processing, we refer interested readers to recent relevant survey papers (Lin et al. 2021; Tay et al. 2020).

Generative models have been attaining much interest in the audio community due to their abilities to learn the underlying audio distribution. Generative adversarial networks (GANs) (Goodfellow et al. 2014), variational autoencoders (VAEs) (Kingma and Welling 2013), and autoregressive models (Shannon et al. 2012) are extensively investigated in the speech and audio processing scientific community. Specifically, they are used to synthesised audio signal from a low-dimensional representation to a high-resolution signal (Hsu et al. 2017; Ma et al. 2019; Latif et al. 2018). The synthesised samples are often used to augment the training material to improve the performance (Latif et al. 2020). In the autoregressive approach, the new samples are synthesised iteratively—based on an infinitely long context of previous samples via RNNs (for example, using LSTM or GRU networks)—but at the cost of expensive computation during training (Wang et al. 2018).

2.2 Reinforcement learning

Reinforcement learning (RL) is a popular paradigm of ML, which involves agents to learn their behaviour by trial and error (Sutton et al. 1998). RL agents aim to learn sequential decision-making by successfully interacting with the environment where they operate. At time t (0 at the beginning of the interaction, T at the end of an episodic interaction, or ∞ in the case of non-episodic tasks), an RL agent in state s_t takes an action $a \in A$, transits to a new state s_{t+1}, and receives reward r_{t+1} for having chosen action a. This process—repeated iteratively—is illustrated in Fig. 3.

An RL agent aims to learn the best sequence of actions, known as *policy*, to obtain the highest overall cumulative reward in the task (or set of tasks) that is being trained on. While it can choose any action from a set of available actions, the set of actions that an agent takes from start to finish is called an *episode*. A Markov decision process (MDP) (Bellman 1966) can be used to capture the episodic dynamics of an RL problem. An MDP can be represented using the tuple (S, A, γ, P, R). The decision-maker or agent chooses an action $a \in A$ in state $s \in S$ at time t according to its policy $\pi(a_t|s_t)$—which determines the agent's way of behaving. The probability of moving to the next state $s_{t+1} \in S$ is given by the state transition function $P(s_{t+1}|s_t, a_t)$. The environment produces a reward $R(s_t, a_t, s_{t+1})$ based on the action taken by the agent at time t. This process continues until the maximum time step or the agent reaches a terminal state. The objective is to maximise the expected discounted cumulative reward, which is given by:

$$E_\pi[R_t] = E_\pi\left[\sum_{i=0}^{\infty} \gamma^i r_{t+i}\right] \tag{1}$$

where $\gamma \in [0,1]$ is a discount factor used to specify that rewards in the distant future are less valuable than in the nearer future. While an RL agent may only learn its policy, it may also learn (online or offline) the transition and reward functions.

3 Deep reinforcement learning

Deep reinforcement learning (DRL) combines conventional RL with DL to overcome the limitations of RL in complex environments with large state spaces or high computation requirements. DRL employs DNNs to estimate value, policy or model that are learnt

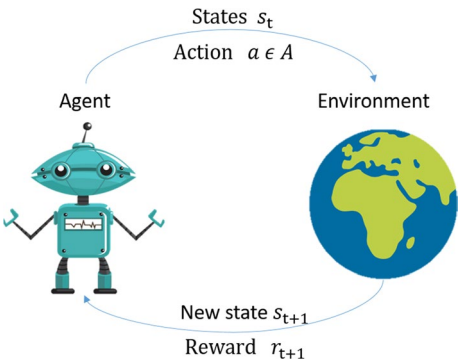

Fig. 3 Basic RL setting

through the storage of state-action pairs in conventional RL (Li 2017). Deep RL algorithms can be classified along several dimensions. For instance, on-policy vs off-policy, model-free vs model-based, value-based vs policy-based DRL algorithms, among others. The salient features of various key characteristics of DRL algorithms are presented and depicted in Fig. 4. Interested readers are referred to Li (2017) for more details on these algorithms. This section focuses on popular DRL algorithms employed in audio-based applications in three main categories: (i) value-based DRL, (ii) policy gradient-based DRL and (iii) model-based DRL.

3.1 Value-based DRL

One of the most famous value-based DRL algorithms is deep Q-network (DQN), introduced by Mnih et al. (2015), that learns directly from high-dimensional inputs. It employs CNNs referred to as Q-network to estimate a value function $Q(s, a)$ by minimizing the loss function at i^{th} iteration given by

$$L_i(\theta_i) = \mathbb{E}_{s,a\sim p(.)}[(y_i - Q(s, a;\theta_i))^2], \tag{2}$$

where $y_i = \mathbb{E}_{s\prime\sim s}[r + \gamma \max_{a\prime} Q(s\prime, a\prime;\theta_{i-1}|s, a]$ defines the i^{th} iteration target and θ represents the weights of the Q-network. DQN enhances data efficiency and stability of the learning process using a technique known as experience replay, where the agent's experience at each time step t, $e_t = \{s_t, a_t, r_t, s_{t+1}\}$ is stored in a replay memory. Subsequently, mini-batches of experience $e \sim D$, where $D = \{e_1, e_2, e_3, \ldots, e_N\}$ are randomly selected and updated using Q-learning. Post-experience replay, the agent applies ϵ-greedy policy to select and execute an action. Although DQN, since inception, has rendered super-human performance in Atari games, it is based on a single max operator, given in (2), for selection as well as evaluation of an action. Thus, the selection of an overestimated action may lead to over-optimistic action value estimates that induces an upward bias. Double DQN (DDQN) (Van Hasseltet et al. 2016) eliminates this positive bias by introducing two

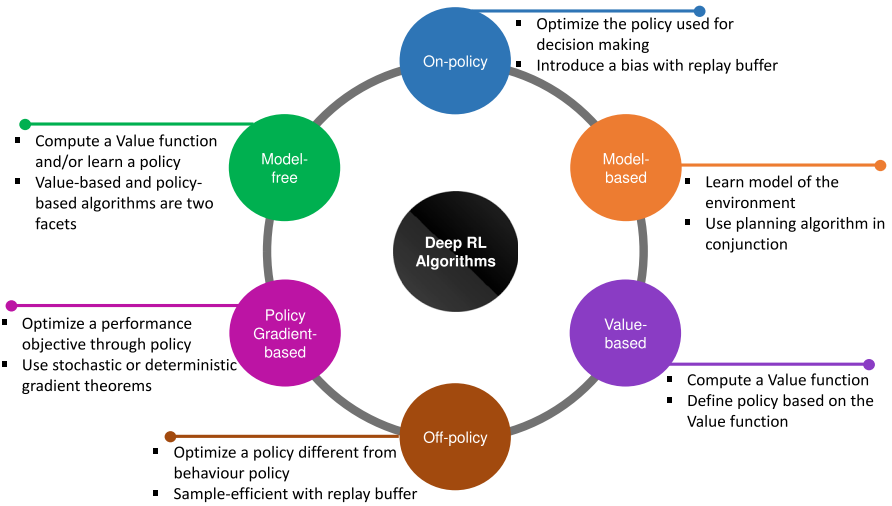

Fig. 4 Characteristics of different DRL algorithms

decoupled estimators: one for the selection of an action, and one for the evaluation of an action. Schaul et al. in Schaul et al. (2016) show that the performance of DQN and DDQN is enhanced considerably if significant experience transitions are prioritised and replayed more frequently. Wang et al. (2016) present a duelling network architecture (DNA) to estimate a value function $V(s)$ and associated advantage function $A(s, a)$ separately, and then combine them to get action-value function $Q(s, a)$. Contrary to DQN that follows convolutional layers with a single fully connected layer, DNA employs two fully connected layers for the estimation of scalar $V(s;\theta, \beta)$ and $|A|$-dimensional vector $A(s, a;\theta, \alpha)$. Here, θ represents convolutional layers parameters, whereas α and β are parameters of two fully connected streams (layers). The two fully connected layers are combined to output a single Q value as per the equation articulated below. Results show that DQN and DDQN having DNA and prioritised experience replay can lead to improved performance.

$$Q(s, a;\theta, \alpha, \beta) = V(s;\theta, \beta) + \left(A(s, a;\theta, \alpha) - \frac{1}{|A|}\sum A(s, a';\theta, \alpha)\right), \tag{3}$$

Unlike the aforementioned DQN algorithms that focus on the expected return, distributional DQN (Bellemare et al. 2017) aims to learn the full distribution of the value in order to have additional information about rewards. Despite both DQN and distributional DQN focusing on maximising the expected return, the latter comparatively results in performant learning. Dabney et al. (2018) propose distributional DQN with quantile regression (QR-DQN) to explicitly model the distribution of the value function. Results demonstrate that QR-DQN successfully bridges the gap between theoretic and algorithmic results. Implicit quantile networks (IQN) (Dabney et al. 2018), an extension to QR-DQN, estimate quantile regression by learning the full quantile function instead of focusing on a discrete number of quantiles. IQN also provides flexibility regarding its training with the required number of samples per update, ranging from one sample to a maximum computationally allowed. IQN has shown to outperform QR-DQN comprehensively in the Atari domain.

The astounding success of DQN to learn rich representations is highly attributed to DNNs, while batch algorithms prove to have better stability and data efficiency (requiring less tuning of hyperparameters). The authors in Levine et al. (2017) propose a hybrid approach named as least-squares DQN (LS-DQN) that exploits the advantages of both DQN and batch algorithms. Deep Q-learning from demonstrations (DQfD) (Hester et al. 2018) leverages human demonstrations to learn at an accelerated rate from the start. Deep quality-value (DQV) (Sabatelli et al. 2018) is a novel temporal-difference-based algorithm that trains the Value network initially, and subsequently uses it to train a Quality-value neural network for estimating a value function. Results in the Atari domain indicate that DQV outperforms DQN as well as DDQN. The authors in Arjona-Medina et al. (2019) propose RUDDER (return decomposition for delayed rewards), which encompasses reward redistribution and return decomposition for Markov decision processes (MDPs) with delayed rewards. Pohlen et al. (2018) employ a transformed Bellman operator along with human demonstrations in the proposed algorithm Ape-X DQfD to attain human-level performance over a wide range of games. Results show that the proposed algorithm achieves average-human performance in 40 out of 42 Atari games with the same set of hyperparameters. Schulman et al. in Schulman et al. (2017) study the connection between Q-learning and policy gradient methods. They show that soft Q-learning (an entropy-regularised version of Q-learning) is equivalent to policy gradient methods and that they perform as well (if not better) than standard variants.

Previous studies have also attempted to incorporate a memory element into DRL algorithms. For instance, the deep recurrent Q-network (DRQN) approach introduced by Hausknecht and Stone (2015) was able to successfully integrate information through time, which performed well on standard Atari games. A further improvement was made by introducing an attention mechanism to DQN, resulting in a deep recurrent Q-network (DARQN) (Sorokin et al. 2015). This allows DARQN to focus on a specific part of the input and achieve better performance compared to DQN and DRQN on games. Some other studies (Oh et al. 2016; Parisotto and Salakhutdinov 2018) have also proposed methods to incorporate memory into DRL, but this area remains to be investigated further.

3.2 Policy gradient-based DRL

Policy gradient-based DRL algorithms aim to learn an optimal policy that maximises performance objectives, such as expected cumulative reward. This class of algorithms make use of gradient theorems to reach optimal policy parameters. Policy gradient typically requires the estimation of a value function based on the current policy. This may be accomplished using the actor-critic architecture, where the *actor* represents the policy and the *critic* refers to value function estimate (Konda and Tsitsiklis 1999). Mnih et al. (Mnih et al. 2016) proposed the advantage actor-critic (A2C) algorithm, which employs an advantage function instead of a value function for updating network weights. The advantage function $A(s_t, a_t) = Q(s_t, a_t) - V(s_t)$ estimates the benefit of a chosen action over an average action for a given state. The authors of Mnih et al. (2016) demonstrate that actor-critic methods yield superior results over value-based methods in terms of training speed. They further show that asynchronous execution of multiple parallel agents on standard CPU-based hardware leads to time-efficient and resource-efficient learning. The proposed asynchronous version of actor-critic, asynchronous advantage actor-critic (A3C) updates policy and value functions after every t_{max} actions or in case a terminal state is reached. For a single update, the agent first receives n-step returns by selecting actions based on its exploration policy till t_{max} steps or terminal state. Afterwards, n-step Q-learning updates are computed for every state-action pair that are further used in the calculation of a single gradient step. A3C exhibits remarkable learning in both 2D and 3D games with action spaces in discrete as well as continuous domains. The authors in Babaeizadeh et al. (2017) propose a hybrid CPU/GPU-based A3C —named as GA3C — showing significantly higher speeds as compared to its CPU-based counterpart.

Asynchronous actor-critic algorithms, including A3C and GA3C, may suffer from inconsistent and asynchronous parameter updates. A novel framework for asynchronous algorithms is proposed in Alfredo et al. (2017) to leverage parallelisation while providing synchronous parameters updates. The authors show that the proposed parallel advantage actor-critic (PAAC) algorithm enables true on-policy learning in addition to faster convergence. The authors in O'Donoghue et al. (2016) propose a hybrid policy-gradient-and-Q-learning (PGQL) algorithm that combines on-policy policy gradient with off-policy Q-learning. Results demonstrate PGQL's superior performance on Atari games as compared to both A3C and Q-learning. Munos et al. (2016) propose a novel algorithm by bringing together three off-policy algorithms: Instance Sampling (IS), Q(λ), and Tree-Backup TB(λ). This algorithm—called Retrace(λ)—alleviates the weaknesses of all three algorithms (IS has low variance, Q(λ) is not safe, and TB(λ) is inefficient) and promises safety, efficiency and guaranteed convergence. Reactor (Retrace-Actor) (Gruslys et al. 2017) is a Retrace-based actor-critic agent architecture

that combines time efficiency of asynchronous algorithms with sample efficiency of off-policy experience replay-based algorithms. Results in the Atari domain indicate that the proposed algorithm performs comparably with state-of-the-art algorithms while yielding substantial gains in terms of training time. The importance of weighted actor-learner architecture (IMPALA) (Espeholt et al. 2018) is a scalable distributed agent that is capable of handling multiple tasks with a single set of parameters. Results show that IMPALA outperforms A3C baselines in a diverse multi-task environment.

Schulman et al. (2015) propose a robust and scalable trust region policy optimisation (TRPO) algorithm for optimising stochastic control policies. TRPO promises guaranteed monotonic improvement regarding the optimisation of nonlinear and complex policies having an inundated number of parameters. This learning algorithm makes use of a fixed KL divergence constraint rather than a fixed penalty coefficient and outperforms a number of gradient-free and policy-gradient methods over a wide variety of tasks. Schulman et al. (2017) introduce proximal policy optimisation (PPO), which aims to be as reliable and stable as TRPO but relatively better in terms of implementation and sample complexity.

3.3 Model-based DRL

Model-based DRL algorithms rely on models of the environment (i.e. underlying dynamics and reward functions) in conjunction with a planning algorithm. Unlike model-free DRL methods that typically entail a large number of samples to render adequate performance, model-based algorithms generally lead to improved sample and time efficiency (Ravindran 2019).

Kaiser et al. (2019) propose simulated policy learning (SimPLe), a video prediction-based model-based DRL algorithm that requires much fewer agent-environment interactions than model-free algorithms. Experimental results indicate that SimPLe outperforms state-of-the-art model-free algorithms in Atari games. Whiteson (2018) propose TreeQN for complex environments, where the transition model is not explicitly given. The proposed algorithm combines model-free and model-based approaches in order to estimate Q-values based on a dynamic tree constructed recursively through an implicit transition model. Authors of Whiteson (2018) also propose an actor-critic variant named ATreeC that augments TreeQN with a softmax layer to form a stochastic policy network. They show that both algorithms yield superior performance than n-step DQN and value prediction networks (Oh et al. 2017) on multiple Atari games. Vezhnevets et al. (2016) introduce a Strategic Attentive Writer (STRAW), which is capable of making natural decisions by learning macro-actions. Unlike state-of-the-art DRL algorithms that yield only one action after every observation, STRAW generates a sequence of actions, thus leading to structured exploration. Experimental results indicate a significant improvement in Atari games with STRAW. Value Propagation (VProp) (Nardelli et al. 2018) is a set of Value Iteration-based planning modules trained using RL and capable of solving unseen tasks and navigating in complex environments. It is also demonstrated that VProp is able to generalise in a dynamic and noisy environment. Schrittwieser et al. (2019) present a model-based algorithm named MuZero that combines tree-based search with a learned model to render superhuman performance in challenging environments. Experimental results demonstrate that MuZero delivers state-of-the-art performance on 57 diverse Atari games. Table 2 presents an overview of DRL algorithms at a glance.

Table 2 Summary of DRL algorithms

DRL algorithms	Approach	Details	Off-policy/on policy
Value-based DRL			
DQN [2]	Target Q-network, experience replay	• Learns directly from high dimensional inputs • Stabilises learning process with target Q-net • Experience replay to avoid divergence	off-policy
DDQN [3]	Double Q-learning	Decoupled estimators for the selection and evaluation of an action	
Prioritised DQN [4]	Prioritised experience replay	Significant experience transitions are prioritised and replayed frequently thus leading to efficient learning	
DNA [5]	Duelling neural network architecture	Estimates a value function and associated advantage function and combine them to get a value function with faster convergence than Q-learning	
Distributional DQN [6]	Learns distribution of cumulative returns using a distributional Bellman equation	• Leads to performant learning than DQN • Possibility to implement risk-aware behaviour	
QR-DQN [7]	Distributional DQN with quantile regression	Bridges the gap between theoretical and algorithmic results	
IQN [8]	Extends QR-DQN with a full quantile function	Provides flexibility regarding number of samples required for training	
LS-DQN [9]	A hybrid approach combining DQN with least-squares method	Exploits advantages of both DQN, ability to learn rich representations, and batch algorithms, stability and data efficiency	
DQfD [10]	Learns from demonstrations	Learns at an accelerated rate from the start.	
DQV [11]	Uses temporal difference to train a Value network and uses it for training a Quality-Value network that estimates state-action values	Learns significantly better and faster than DQN and DDQN.	
RUDDER [12]	Reward redistribution and return decomposition.	Provides prominent improvement on games having long delayed rewards	

Table 2 (continued)

DRL algorithms	Approach	Details	Off-policy/on policy
Value-based DRL			
Ape-X DQfD [13]	Employs a transformed Bellman operator together with a temporal consistency loss	Surpasses average human performance on 40 out of 42 Atari 2600 games	
Soft DQN [14]	Incorporation of soft KL penalty and entropy bonus	Establishes equivalence between Soft DQN and policy gradient	
DRQN, DARQN [14-15]	Memory, attention	DQN policies modelled by attention-based recurrent networks.	
Policy Gradient-based DRL			
A3C [16]	Asynchronous gradient descent	Consumes less resources; able to run on a standard multi-core CPU	on-policy
GA3C [17]	Hybrid CPU/GPU-based A3C	Achieves speed significantly higher than its CPU-based counterpart	
PAAC [18]	Novel framework for asynchronous algorithms	Computationally efficient & enables faster convergence to optimal policies	
PGQL [19]	Combines on-policy policy gradient with off-policy Q-learning	Enhanced stability and data efficiency	off-policy
Retrace(λ) [20]	Expresses 3 algorithms in a common form: IS, Q(λ) and TB(λ)	Safe, sample efficient and has low variance	
Reactor [21]	Retrace-based actor-critic agent architecture	Yields substantial gains in terms of training time.	
IMPALA [22]	Scalable distributed agent capable of handling multiple tasks with a single set of parameters	outperforms state-of-the-art agents in a diverse multi-task environment	on-policy
TRPO [23]	Employs fixed KL divergence constraint for optimising stochastic control policies	Performs well over a wide variety of large-scale tasks	

Table 2 (continued)

DRL algorithms	Approach	Details	Off-policy/on policy
Value-based DRL			
PPO [24]	Makes use of an adaptive KL penalty coefficient	As reliable and stable as TRPO but relatively better in terms of implementation and sample complexity	
Model-based DRL			
SimPLe [26]	Video prediction model-based algorithm requiring much fewer agent-environment interactions than model-free algorithms	Outperforms state-of-the-art model-free algorithms in Atari games	on-policy
TreeQN [27]	Estimates Q-values using a dynamic tree built recursively with a transition model	Outperforms n-step DQN and value prediction networks in multiple Atari games	
STRAW [29]	Capable of decision making by learning macro actions	Improves performance significantly in Atari games	
VProp [30]	Value Iteration planning modules trained with RL	• Solves an unseen task and navigate in complex environments • Generalise in dynamic & noisy environments	–
MuZero [31]	Combines tree-based search with a learned model	Delivers state-of-the-art performance on 57 diverse Atari games	Off-policy

3.4 Audio processing using DRL

Audio processing using DRL include different components including environment, agent, action, and reward. Audio is a 1-dimensional (1D) time-series signal that goes through different pre-processing and feature extraction procedures. Pre-processing steps involve noise suppression, silence removal, and channel equalisation, which enhances audio signal quality to build robust and efficient audio-based systems. It has been found that pre-processing helps to improve DL-based audio systems (Latif et al. 2020). Feature extraction usually comes after pre-processing, which aims to convert an audio signal into meaningful, informative, and a reasonably limited number of features. Mel-frequency cepstral coefficients (MFCCs) and spectrograms are considered a popular choice of input features in audio-based systems (Latif et al. 2020). These features are given to the DRL agent to perform different tasks based on the application. An example scenario is a human speaking to a machine trained via DRL as in Fig. 5, where the machine has to act based on features derived from audio (among other) signals. We discuss in detail different audio-based systems next.

4 Audio-based DRL

This section surveys related works where audio is a key element in the learning environments of DRL agents. Table 3 summarises the characterisation of DRL agents for six audio-related areas: (I) automatic speech recognition; (II) spoken dialogue systems; (III) emotions modelling; (IV) audio enhancement; (V) music listening and generation; and (VI) human–robot interaction (HRI). There is a large literature on audio-based DRL and it is used in a wide variety of applications. Therefore and in order to keep this review to a manageable length, we limit ourselves to these six main areas here. In Sect. 4.7 we briefly mention some remaining audio-related areas and other applications.

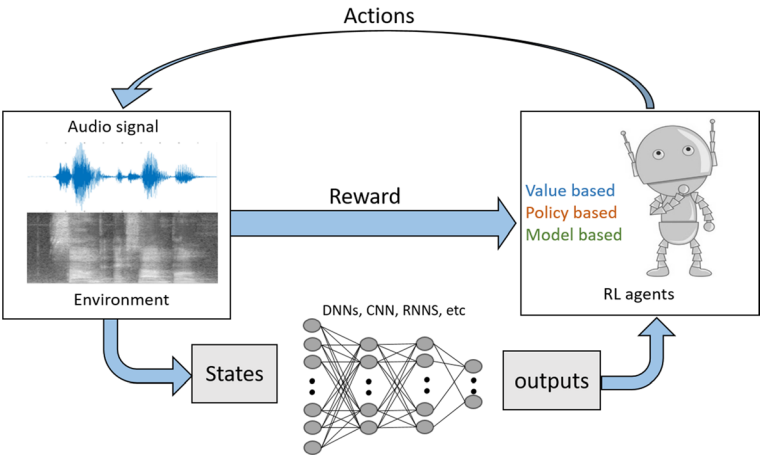

Fig. 5 Schematic diagram of DRL agents for audio-based applications, where the DL model (via DNNs, CNN, RNNs, etc.) generates audio features from raw waveforms or other audio representations for taking actions that change the environment from state s_t to a next state s_{t+1}

Table 3 Summary of audio related fields, characterisation of DRL agents, and related datasets

Area	State Representations $\mathcal{S}$, Actions $\mathcal{A}$, and Reward Functions $\mathcal{R}$	Popular Datasets
Automatic Speech Recognition	$\mathcal{S}$: States are learnt features from input speech features (fMLLR/MFCC vectors (Rath et al. 2013)). $\mathcal{A}$: They include phonemes, graphemes, commands, or candidates from ASR results. $\mathcal{R}$: They have included binary rewards (positive for correct choices, 0 otherwise), and non-binary sentence/token rewards based on the Levenshtein distance algorithm.	• LibriSpeech (Panayotov et al. 2015) • TED-LIUM (Rousseau et al. 2012) • Wall Street Journal (Paul and Baker 1992) • SWITCHBOARD (Godfrey et al. 1992) • TIMIT (Garofolo et al. 1993)
Spoken Dialogue Systems	$\mathcal{S}$: They encode uttered words by the system and recognised user words into a dialogue history. Additional information includes classification results such as user goals, user intents, speech recognition confidence scores, and visual information (in the case of multimodal systems), among others. $\mathcal{A}$: While in task-oriented systems they include slot requests/confirmations/apologies, slot-value selection, ask question, data retrieval, information presentation, among others; actions in open-ended systems include either all possible sentences (infinite) or clusters of sentences (finite). $\mathcal{R}$: They vary depending on the company/project requirements and tend to include sparse and non-sparse numerical rewards such as dialogue length, task success, dialogue similarity, dialogue coherence, dialogue repetitiveness, game scores (in the case of game-based systems), among others—see more details in Table 4.	• SGD (Rastogi et al. 2020) • DSTC (Williams et al. 2016) • Frames (Asri et al. 2017) • MultiWOZ (Budzianowski et al. 2018) • SubTle Corpus (Ameixa et al. 2013) • Simulations (Schatzmann et al. 2006) • Other datasets (Serban et al. 2018)
Speech Emotion Recognition	$\mathcal{S}$: Speech features (e.g., MFCC) are considered input features. $\mathcal{A}$: Actions include speech emotion labels (e.g. unhappy, neutral, happy), sentiment detection (e.g. negative, neutral, positive), and termination from utterance listening. $\mathcal{R}$: Binary rewards have been used (positive for correct choices, 0 otherwise).	• EMODB (Burkhardt et al. 2005) • IEMOCAP (Busso et al. 2008) • MSP-IMPROV (Busso et al. 2016) • SEMAINE (McKeown et al. 2011) • MELD (Poria et al. 2019)
Audio Enhancement	$\mathcal{S}$: States are learnt from clean and noisy acoustic features. $\mathcal{A}$: Finding closest cluster and its index, time-frequency mask estimation, and increasing or decreasing the parameter values of the speech-enhancement algorithm. $\mathcal{R}$: Positive rewards for the correct choice, negative otherwise.	• DEMAND (Thiemann et al. 2013) • CHiME-3 (Barker et al. 2015) • WHAMR (Maciejewski et al. 2020)
Music Generation	$\mathcal{S}$: State representations are learned from Musical notes. $\mathcal{A}$: Musical generation and next note selection are considered actions. $\mathcal{R}$: Binary reward functions based on hard-coded musical theory rules, including the likelihood of actions.	• Classical piano MIDI database (Krueger 2016) • MusicNet dataset (Thickstun et al. 2016) • JSB Chorales dataset (Allan and Williams 2005)

Table 3 (continued)

Area	State Representations $\mathcal{S}$, Actions $\mathcal{A}$, and Reward Functions $\mathcal{R}$	Popular Datasets
Robotics, Control & Interaction	$\mathcal{S}$: They encode visual and verbal representations derived from image embeddings, speech features, and word or sentence embeddings. Additional information includes user intents, speech recognition scores, human activities, postures, emotions, and body joint angles, among others. $\mathcal{A}$: They include motor commands (e.g. gestures, locomotion, navigation, manipulation, gaze) and verbalizations such as dialogue acts and backchannels (e.g. laughs, smiles, noddings, head-shakes). $\mathcal{R}$: They are based on task success (positive rewards for achieving the goal, negative rewards for failing the task, and zero/shaped rewards otherwise) and user engagement.	• AVDIAR (Gebru et al. 2017) • NLI Corpus (Scalise et al. 2018) • VEIL dataset (Misra et al. 2016) • Simulations (Cuayáhuitl 2020) • Real-world interactions (Qureshi et al. 2018)

4.1 Automatic speech recognition (ASR)

Automatic speech recognition (ASR) is the process of converting a speech signal into its corresponding text by using algorithms. Contemporary ASR technology has reached great levels of performance due to advancements in DL techniques. The performance of ASR systems, however, relies heavily on supervised training of deep models with large amounts of transcribed data. Even for resource-rich languages, additional transcription costs required for new tasks hinders the applications of ASR. To broaden the scope of ASR, different studies have attempted DRL based models to learn from feedback or environment. This form of learning aims to reduce transcription costs and time by providing positive or negative rewards instead of detailed transcriptions. For instance, Kala and Shinozaki (2018) proposed an RL framework for ASR based on the policy gradient method that provides a new view of existing training and adaptation methods. This makes the ASR system self-sufficient to learn from feedback of users and help achieve improved speech recognition performance and reduced Word Error Rate (WER) compared to unsupervised adaptation. In ASR, sequence-to-sequence models have shown great success; however, these models fail to approximate real-world speech during inference. Tjandra et al. (2018) solved this issue by training a sequence-to-sequence model with a policy gradient algorithm. In contrast to standard training on maximum likelihood estimation (MLE), they used policy gradient to sample the whole transcription by directly optimising the negative Levenshtein distance as the reward. Their results showed a significant improvement using an RL-based objective and an MLE objective compared to the model trained with only the MLE objective. In another study, (Tjandra et al. 2019) the authors found that using token-level rewards (intermediate rewards are given after each time step) provides improved performance compared to sentence-level rewards and baseline systems. In order to solve the issues of semi-supervised training of sequence-to-sequence ASR models, Chung et al. (2020) investigated the REINFORCE algorithm by rewarding the ASR to output more correct sentences for both unpaired and paired speech input data. Experimental evaluations showed that the DRL-based method was able to effectively reduce character error rates from 10.4 to 8.7%.

Karita et al. (2018) propose to train an encoder-decoder ASR system using a sequence-level evaluation metric based on the policy gradient objective function. This enables the minimisation of the expected WER of the model predictions. In this way, the authors found that the proposed method improves speech recognition performance. The ASR system of Zhou et al. (2018) was jointly trained with maximum likelihood and policy gradient to improve via end-to-end learning. They were able to optimise the performance metric directly with policy learning and achieve 4% to 13% relative improvement for end-to-end ASR. In Luo et al. (2017), the authors attempted to solve sequence-to-sequence problems by proposing a model based on supervised backpropagation and a policy gradient method, which can directly maximise the log probability of the correct answer. They achieved very encouraging results on a small scale and a medium scale ASR. Radzikowski et al. (2019) proposed a dual supervised model based on a policy gradient methodology for non-native speech recognition. They evaluated tested warm-start and semi warm-start approaches, and were able to achieve promising results for the English language pronounced by Japanese and Polish speakers.

To achieve the best possible accuracy, end-to-end ASR systems are becoming increasingly large and complex. DRL methods can also be leveraged to provide model compression (He et al. 2018). In Dudziak et al. (2019), RL-based ShrinkML is proposed

to optimise the per-layer compression ratios in a state-of-the-art LSTM-based ASR model with attention. They exploited RL to push the boundaries of singular value decomposition (SVD) based ASR mode compression. Evaluations were preformed on LibriSpeech data. Based on the results, the authors found that the RL-based model was able to effectively compress a ASR system compared to the manually-compressed models. For time-efficient ASR, Rajapakshe et al. (2020) evaluated the pre-training of an RL-based policy gradient network. They found that pre-training in DRL offers faster convergence compared to non-pre-trained networks, and also achieve improved recognition in lesser time. To tackle the slow convergence time of the REINFORCE algorithm (Williams 1992; Lawson et al. 2018), evaluated Variational Inference for Monte Carlo Objectives (VIMCO) and Neural Variational Inference (NVIL) for phoneme recognition tasks in clean and noisy environments. The authors found that the proposed method (using VIMCO and NVIL) outperforms REINFORCE and other methods at training online sequence-to-sequence models.

The studies above highlight ways to improve the performance of ASRs by involving interaction with the environment using DRL. Despite these promising results, further research is required on DRL algorithms towards building autonomous ASR systems that can work in complex real-life settings. The REINFORCE algorithm is very popular in ASR, therefore, research is also required to explore other DRL algorithms to highlight its suitability for ASR.

4.2 Spoken dialogue systems (SDSs)

Spoken dialogue systems are gaining interest due to many applications in customer services and goal-oriented human-computer interaction. Typical SDSs integrate several key components including speech recogniser, intent recogniser, knowledge base and/or database backend, dialogue manager, language generator, and speech synthesis, among others (Zue and Glass 2000). The task of a dialogue manager in SDSs is to select actions based on observed events (Levin et al. 2000; Singh et al. 2000). Researchers have shown that the action selection process can be effectively optimised using RL to model the dynamics of spoken dialogue as a fully or partially observable Markov Decision Process (Paek 2006). Numerous studies have utilised RL-based algorithms in spoken dialogue systems. In contrast to text-based dialogue systems that can be trained directly using large amounts of text data (Gao et al. 2019), most SDSs have been trained using user simulations (Schatzmann et al. 2006). The justification for that is mainly due to insufficient amounts of training dialogues to train or test from real data (Serban et al. 2018).

SDSs involve policy optimisation to respond to humans by taking the current state of the dialogue, selecting an action, and returning the verbal response of the system. For instance, Chen et al. (Chen et al. 2020) presented an online DRL-based dialogue state tracking framework in order to improve the performance of a dialogue manager. They achieved promising results for online dialogue state tracking in the second and third dialogue state tracking challenges (Henderson et al. 2014, 2014). Weisz et al. (2018) utilised DRL approaches, including actor-critic methods and off-policy RL. They also evaluated actor-critic with experience replay (ACER) (Wang et al. 2016; Munos et al. 2016), which has shown promising results on simple gaming tasks. They showed that the proposed method is sample efficient and that performed better than some state-of-the-art DL approaches for spoken dialogue. A task-oriented end-to-end DRL-based dialogue system is proposed in Cuayáhuitl (2017). They showed that DRL-based optimisation produced significant

improvement in task success rate and also caused a reduction in dialogue length compared to supervised training. Zhao and Eskenazi (2016) utilised deep recurrent Q-networks (DRQN) for dialogue state tracking and management. Experimental results showed that the proposed model can exploit the strengths of DRL and supervised learning to achieve faster learning speed and better results than the modular-based baseline system. To present baseline results, a benchmark study (Casanueva et al. 2017) is performed using DRL algorithms including DQN, A2C and natural actor-critic (Su et al. 2017) and their performance is compared against GP-SARSA (Gašić et al. 2013). Based on experimental results on the PyDial toolkit (Ultes et al. 2017), the authors conclude that substantial improvements are still needed for DRL methods to match the performance of carefully designed handcrafted policies.

In addition to SDSs optimised via flat DRL, hierarchical RL/DRL methods have been proposed for policy learning using dialogue states with different levels of abstraction and dialogue actions at different levels of granularity (via primitive and composite actions) (Cuayáhuitl 2009; Cuayáhuitl et al. 2010; Dethlefs and Cuayáhuitl 2015; Budzianowski et al. 2017; Peng et al. 2017; Zhang et al. 2018). The benefits of this form of learning include faster training and policy reuse. A deep Q-network based multi-domain dialogue system is proposed in Cuayáhuitl et al. (2016). They train the proposed SDS using a network of DQN agents, which is similar to hierarchical DRL but with more flexibility for transitioning across dialogues domains. Another work-related to faster training is proposed by Gordon-Hall et al. (2020), where the behaviour of RL agents is guided by expert demonstrations.

The optimisation of dialogue policies requires a reward function that unfortunately is not easy to specify. Unless a clear and concrete performance function is available (rather unlikely), this stage may require annotated data for training a reward predictor instead of a hand-crafted one. In real-world applications, such annotations are either scarce or not available. Therefore, some researchers have turned their attention to methods for online active reward learning. In Su et al. (2016), the authors presented an online learning framework for a spoken dialogue system. They jointly trained the dialogue policy alongside the reward model via active learning. Based on the results, the authors showed that the proposed framework can significantly reduce data annotation costs and can also mitigate noisy user feedback in dialogue policy learning. Su et al. (2017) introduced two approaches: trust region actor-critic with experience replay (TRACER) and episodic natural actor-critic with experience replay (eNACER) for dialogue policy optimisation. From these two algorithms, they achieved the best performance using TRACER.

In Ultes et al. (2017), the authors propose to learn a domain-independent reward function based on user satisfaction for dialogue policy learning. The authors showed that the proposed framework yields good performance for both task success rate and user satisfaction. Researchers have also used DRL to learn dialogue policies in noisy environments, and some have shown that their proposed models can generate dialogues indistinguishable from human ones (Fazel-Zarandi et al. 2017). Carrara et al. (2017) propose a clustering approach for online user adaptation in RL-based dialogue systems. They propose a distance metric and build on previous works in an attempt to reduce the number of possible transfers from similar users. Experiments were carried out on a negotiation dialogue task, which showed significant improvements over baselines. In another study (Carrara et al. 2018), authors proposed ϵ-safe, a Q-learning algorithm, for safe transfer learning for dialogue applications. A DRL-based chatbot called MILABOT was designed in Serban et al. (2017), which can converse with humans on popular topics through both speech and text—performing significantly

better than many competing systems. The text-based chatbot in Cuayáhuitl et al. (2019) used an ensemble of DRL agents and showed that training multiple dialogue agents performs better than a single agent.

Table 4 shows a summary of DRL-based (spoken) dialogue systems. While not all involve spoken interactions, they can be applied to speech-based systems by for example using the outputs from speech recognisers instead of typed interactions, i.e. ASR systems can be seen as feature extractors from audio data for dialogic interaction. While at first instance it may look like text-based DRL agents would not be able to cope with noisy inputs, using ASR-based inputs can be useful because they can be enriched with word, sentence, and/or knowledge embeddings for dealing with unobserved utterances during training. This form of generalisation would be hard to achieve (if not impossible) using only audio-based features. In addition, modelling dialogue history taking features from multiple utterances in a conversation can help to deal with noisy inputs. But what features to include in the dialogue history over multiple turns has been and it is still task-specific—task-agnostic features is something that needs to be investigated further to benefit the creation of applications bootstrapped by previous ones. Including low and high-level features ranging from speech features, to multimodal and knowledge features is something that requires further understanding in order to draw recommendations for different applications. As a matter of fact and to our knowledge, the inclusion of audio features in the dialogue state has been overlooked in SDSs (with the exception of Zorrilla et al. (2021)) and it could prove useful, but this remains to be investigated further.

In terms of application, we can observe in Table 4 that most systems focus on one or a few domains (or tasks)—systems trained with a large number of domains is usually not attempted, presumably due to the high requirements of data and compute involved. Regarding algorithms, the most popular are DQN-based or REINFORCE among other more recent algorithms—when to use one over another algorithm still needs to be understood better. We can also observe that user simulations are mostly used for training task-oriented dialogue systems, while real data is the preferred choice for open-ended dialogue systems. We can note that while transfer learning is an important component in a trained SDS, it is not commonplace yet. Given that learning from scratch every time a system is trained is neither scalable nor practical, it looks like transfer learning will naturally be adopted more and more in the future as more domains are taken into account. In terms of datasets, most of them are still small size. It is rare to see SDSs trained with millions of dialogues or sentences. As datasets grow, the need for more efficient training methods will take more relevance in future systems. Regarding human evaluations, we can observe that about half of research works involve human evaluations. While human evaluations may not always be required to answer a research question, they certainly should be used whenever learnt conversational skills are being assessed or judged. We can also note that there is no standard for specifying reward functions due to the wide variety of functions used in previous works—almost every paper uses a different reward function. Even when some works use learnt reward functions (e.g. based on adversarial learning), they focus on learning to discriminate between machine-generated and human-generated dialogues without taking other dimensions into accounts such as task success or additional penalties. Although there is advancement in the specification of reward functions by learning them instead of hand-crafting them, this area requires better understanding for optimising different types of dialogues including information-seeking, chitchat, game-based, negotiation-based, etc.

4.3 Emotions modelling

Emotions are essential in vocal human communication, and they have recently received growing interest by the research community (Latif et al. 2019; Wang et al. 2020; Latif 2020; Ali et al. 2021). Arguably, human–robot interaction can be significantly enhanced if dialogue agents can perceive the emotional state of a user and its dynamics (Ma et al. 2020; Majumder et al. 2019). This line of research is categorised into two areas: emotion recognition in conversations (Poria et al. 2019), and affective dialogue generation (Young et al. 2020; Zhou et al. 2018). Speech emotion recognition (SER) can be used as a reward for RL based dialogue systems (Heusser et al. 2019). This would allow the system to adjust the behaviour based on the emotional states of the dialogue partner. Lack of labelled emotional corpora and low accuracy in SER are two major challenges in the field. To achieve the best possible accuracy, various DL-based methods have been applied to SER, however, performance improvement is still needed for real-time deployments. DRL offers different advantages to SER, as highlighted in different studies. In order to improve audio-visual SER performance, Ouyang et al. (2018) presented a model-based RL framework that utilised feedback of testing results as rewards from the environment to update the fusion weights. They evaluated the proposed model on the Multimodal Emotion Recognition Challenge (MEC 2017) dataset and achieved top 2 at the MEC 2017 Audio-Visual Challenge. To minimise the latency in SER, Lakomkin et al. (2018) proposed EmoRL for predicting the emotional state of a speaker as soon as it gains enough confidence while listening. In this way, EmoRL was able to achieve lower latency and minimise the need for audio segmentation required in DL-based approaches for SER. In Sangeetha and Jayasankar (2019), authors used RL with an adaptive fractional deep Belief network (AFDBN) for SER to enhance human-computer interaction. They showed that the combination of RL with AFDBN is efficient in terms of processing time and SER performance. Another study (Chen et al. 2017) utilised an LSTM-based gated multimodal embedding with temporal attention for sentiment analysis. They exploited the policy gradient method REINFORCE to balance exploration and optimisation by random sampling. They empirically show that the proposed model was able to deal with various challenges of understanding communication dynamics.

DRL is less popular in SER compared ASR and SDSs. The above-mentioned studies attempted to help solve different SER challenges using DRL, however, there is still a need for developing adaptive SER agents that can perform SER in the wild using small amounts (few) of samples of data (Latif et al. 2020; Ntalampiras 2021; Latif et al. 2020, 2021). Researchers have been motivated in exploring transfer learning in SER (Ntalampiras 2017) to utilise external knowledge for accelerating the learning process of agents.

4.4 Audio enhancement

The performance of audio-based intelligent systems is critically vulnerable to noisy conditions and degrades according to the noise levels in the environment (Li et al. 2015). Several approaches have been proposed (Latif et al. 2018; Li et al. 2013) to address problems caused by environmental noise. One popular approach is audio enhancement, which aims to generate an enhanced audio signal from its noisy or corrupted version (Wang and Wang 2016). DL-based speech enhancement has attained increased attention due to its superior performance compared to traditional methods (Baby et al. 2015; Wang and Chen 2018).

In DL-based systems, the audio enhancement module is generally optimised separately from the main task such as minimisation of WER. Besides the speech enhancement module, there are different other units in speech-based systems which increase their complexity and make them non-differentiable. In such situations, DRL can achieve complex goals in an iterative manner, which makes it suitable for such applications. Such DRL-based approaches have been proposed in Shen et al. (2019) to optimise the speech enhancement module based on the speech recognition results. Experimental results have shown that DRL-based methods can effectively improve the system's performance by 12.4% and 19.2% error rate reductions for the signal to noise ratio at 0 dB and 5 dB, respectively. In Koizumi et al. (2017), authors attempted to optimise DNN-based source enhancement using RL with numerical rewards calculated from conventional perceptual scores such as perceptual evaluation of speech quality (PESQ) (Recommendation 2001) and perceptual evaluation methods for audio source separation (PEASS) (Emiya et al. 2011). They showed empirically that the proposed method can improve the quality of the output speech signals by using RL-based optimisation. Fakoor et al. (2017) performed a study in an attempt to improve the adaptivity of speech enhancement methods via RL. They propose to model the noise-suppression module as a black box, requiring no knowledge of the algorithmic mechanics. Using an LSTM-based agent, they showed that their method improves system performance compared to methods with no adaptivity. In Alamdari et al. (2020), the authors presented a DRL-based method to achieve personalised compression from noisy speech for a specific user in a hearing aid application. To deal with non-linearities of human hearing via the reward/punishment mechanism, they used a DRL agent that receives preference feedback from the target user. Experimental results showed that the developed approach achieved preferred hearing outcomes.

Similarly to SER, very few studies have explored DRL for audio enhancement. Most of these studies have evaluated DRL-based methods to achieve a certain level of signal enhancement in controlled environments. Further research efforts are needed to develop DRL agents that can perform their tasks in real and complex noisy environments.

4.5 Music listening and generation

DL models are widely used for generating content including images, text, and music. The motivation for using DL for music generation lies in its generality since it can learn from arbitrary corpora of music and be able to generate various musical genres compared to classical methods (Steedman 1984; Ebcioğlu 1988).

Here, DRL offers opportunities to impose rules of music theory for the generation of more real musical structures (Jaques et al. 2016). Various researchers have explored such opportunities of DRL for music generation. For instance, Kotecha (2018) explored DQN to impose greater global coherence and encourage exploration in music generation. Based on the evaluations, the authors achieved better quantitative and qualitative results using an LSTM-based architecture in generating polyphonic music aligned with musical rules. Jiang et al. (2020) presented an interactive RL-Duet framework for real-time human-machine duet improvisation. They used actor-critic with generalised advantage estimator (GAE) (Schulman et al. 2016) based music generation agent to learn a policy for generating musical note generation based on the previous context. They trained the model on monophonic and polyphonic data and were able to generate high-quality musical pieces compared to a baseline method. Jaques et al. (2016) utilised a deep Q-learning agent with a reward function based on rules of music theory and probabilistic outputs of an RNN. They showed that

the proposed model can learn composition rules while maintaining the important information of data learned from supervised training. For audio-based generative models, it is often important to tune the generated samples towards some domain-specific metrics. To achieve this, Guimaraes et al. (2017) proposed a method that combines adversarial training with RL. Specifically, they extend the training process of a GAN framework to include the domain-specific objectives in addition to the discriminator reward. Experimental results show that the proposed model can generate music while maintaining the information originally learned from data, and attained improvement in the desired metrics. In Lee et al. (2017), the authors also used a GAN-based model for music generation and explored optimisation via RL. They used RNN based generator to learn musical distributions from the embedded space and found that the proposed framework was able to generate musically coherent sequences with improved quantitative and qualitative measures. RaveForce (Lan et al. 2019) is a DRL-based environment for music generation, which can be used to search new synthesis parameters for a specific timbre of an electronic musical note or loop.

Score following is the process of tracking a musical performance for a known symbolic representation (a score). Dorfer et al. (2018) modelled score following task with DRL algorithms such as synchronous advantage actor-critic (A2C). They designed a multimodal RL agent that listens to music, reads the score from an image and follows the audio in an end-to-end fashion. Experiments on monophonic and polyphonic piano music showed promising results compared to state-of-the-art methods. The score following task is studied in Henkel et al. (2019) using the A2C and proximal policy optimisation (PPO). This study showed that the proposed approach could be applied to track real piano recordings of human performances.

4.6 Human robot interaction (HRI)

There is a recent growing research interest in robotics to enable robots with abilities such as recognition of users' gestures and intentions (Howard and Cambria 2013), and generation of socially appropriate speech-based behaviours (Goodrich and Schultz 2007). In such applications, RL is suitable because robots are required to learn from rewards obtained from their actions. Different studies have explored different DRL-based approaches for audio and speech processing in robotics. Gao et al. (2020) simulated an experiment for the acquisition of spoken language to provide a proof-of-concept of Skinner's idea (Skinner et al. 1957), which states that children acquire language based on behaviourist reinforcement principles by associating words with meanings. Based on their results, the authors were able to show that acquiring spoken language is a combination of observing the environment, processing the observation, and grounding the observations with their true meaning through a series of reinforcement attempts. In Yu et al. (2018), authors build a virtual agent for language learning in a maze-like world. It interactively acquires the teacher's language from question answering sentence-directed navigation. Some other studies (Sinha et al. 2019; Hermann et al. 2017; Hill et al. 2018) in this direction have also explored RL-based methods for spoken language learning.

In human–robot interaction, researchers have used audio-driven DRL for robot gaze control and dialogue management. In Lathuilière et al. (2019), the authors used Q-learning with DNNs for audio–visual gaze control with the specific goal of finding good policies to control the orientation of a robot head towards groups of people using audio-visual information. Similarly, the authors of Lathuilière et al. (2018) used a deep Q-network taking into account visual and acoustic observations to direct the robot's head towards targets of

interest. Based on the results, the authors showed that the proposed framework generates state-of-the-art results. Clark-Turner and Begum (2018) proposed an end-to-end learning framework that can induce generalised and high-level rules of human interactions from structured demonstrations. They empirically show that the proposed model was able to identify both auditory and gestural responses correctly. Another interesting work (Hussain et al. 2019) utilised a deep Q-network for speech-driven backchannels like laugh generation to enhance engagement in human–robot interaction. Based on their experiments, they found that the proposed method has the potential of training a robot for engaging behaviours. Similarly, Hussain et al. (2019) utilised recurrent Q-learning for backchannel generation to engage agents during human–robot interaction. They showed that an agent trained using off-policy RL produces more engagement than an agent trained from imitation learning. In a similar strand, Bui and Chong (2019) have applied a deep Q-network to control the speech volume of a humanoid robot in environments with different amounts of noise. In a trial with human subjects, participants rated the proposed DRL-based solution better than fixed-volume robots. DRL has also been applied to spoken language understanding (Zamani et al. 2018), where a deep Q-network receives symbolic representations from an intent recogniser and outputs actions such as `(keep mug on sink)`. In Qureshi et al. (2018), the authors trained a humanoid robot to acquire social skills for tracking and greeting people. In their experiments, the robot learnt its human-like behaviour from experiences in a real uncontrolled environment. In Cuayáhuitl (2020), they propose an approach for efficiently training the behaviour of a robot playing games using a very limited amount of demonstration dialogues. Although the learnt multimodal behaviours are not always perfect (due to noisy perceptions), they were reasonable while the trained robot interacted with real human players. Efficient training has also been explored using interactive feedback from human demonstrators as in Moreira et al. (2020), who show that DRL with interactive feedback leads to faster learning and with fewer mistakes than autonomous DRL (without interactive feedback).

Robotics plays an interesting role in bringing audio-based DRL applications together including all or some of the above. For example, a robot recognising speech and understanding language (Zamani et al. 2018), aware of emotions (Lakomkin et al. 2018), carry out activities such as playing games (Cuayáhuitl 2020), greeting people (Qureshi et al. 2018), or playing music (Fryen et al. 2020), among others. Such a collection of DRL agents are currently trained independently, but we should expect more connectedness between them in the future.

4.7 Other applications

Besides the above-mentioned applications, DRL has also been being explored in various audio-based domains including audio localisation, audio scene analysis, speech synthesis, soundscape, and bio-acoustics. In these domains, we found very few studies that focused on DRL. Speech synthesis, also known as text-to-speech (TTS), is an important audio technology that aims to generate human-like natural-sounding speech using text data as input (Latif et al. 2021). Most of the neural speech synthesis systems utilise linguistic or acoustic features as an intermediate representation to generate speech. In the speech synthesis domain, deep end-to-end models (e.g., Shen et al. 2018; Łańcucki et al. 2021; Ren et al. 2019) have attained considerable attention by significantly enhancing the quality of synthesised speech (Latif et al. 2021). Recently, some studies have explored DRL for TTS. For instance, Liu et al. (2021) used RL for emotional speech synthesis. The authors focused

on the use of RL to solve the problem of emotion confusion in TTS systems via interaction between the model and SER. In their experiments, they found that the proposed framework outperformed the state-of-the-art baselines by improving the emotion discriminability of synthesised speech. Mohan et al. (2020) used RL for deciding interleaved actions in sequence-to-sequence models for incremental TTS. Based on their results, the authors found that RL agents can successfully balance the trade-off between the quality of the synthesised speech against the latency of generation. Chung et al. (2021) presented a Reinforce-Aligner—an RL based alignment search agent that can perform optimal duration predictions based on the actions and cumulative rewards. Results showed that the proposed framework can perform accurate alignments of phoneme-to-frame sequence, which help improve the naturalness and fidelity of synthesised speech. DRL has also been applied to bioacoustics (Ntalampiras 2018) and sound emotion (Huang et al. 2019), which need further research.

Research studies have also explored the potentials of DRL for solving audio localisation problems. For instance, Giannakopoulos et al. (2021) trained an autonomous agent that navigates in a two-dimensional space using audio information from a multi-speaker environment. Based on their results, the authors found that the agent can successfully localise the target speaker among a set of predefined speakers in a room by avoiding confusion and going outside the predefined room boundaries. Self-supervised learning (SSL) has been actively studied recently in various research fields including audio, text, vision, and many more (Xin et al. 2020; Latif et al. 2021, 2022). (Gonzalez-Billandon et al. 2020) used a self-supervised RL-based iCub humanoid robot for speaker localisation in an autonomous way. During experimentation, they created a dataset of audio and location mapping which can be utilised to train an agent for accurate and robust speaker localisation. Seurin et al. (2020) presented an RL-based interactive speaker recognition system that aims to improve its performance by requesting personalised utterances to learn speaker representations. They empirically showed that the proposed architecture improves speaker identification compared to the non-interactive baseline models. Shah (Shah et al. 2018) et al. presented FollowNet, a DRL agent that navigates following natural language directions. They empirically showed that FollowNet can successfully navigate by learning to execute previously unseen instructions with a 30% improvement in results over a baseline. Some other studies have also exploited DRL methods for audio-visual navigation (Chen et al. 2020, 2019; Gan et al. 2020) and source separation (Majumder et al. 2021).

In reinforcement learning applications, defining an appropriate reward function to achieve the desired behaviour is challenging. Inverse reinforcement learning (IRL) facilities an automatic way of finding a reward function based on the given set of trajectories in the environment (Ng et al. 2000; Abbeel and Ng 2004). A few studies have explored IRL to impose a learnt reward function, instead manually defined, for dialogue control (Sugiyama et al. 2012) and interactive systems. However, further research efforts are required in the audio domain for designing optimal reward functions.

5 Challenges in audio-based DRL

The research works in the previous section have focused on a narrow set of DRL algorithms and have ignored the existence of many other algorithms, as can be noted in Fig. 6. This suggests the need for a stronger collaboration between core DRL and audio-based DRL, which may be already happening. Figure 7 help us to illustrate that previous works

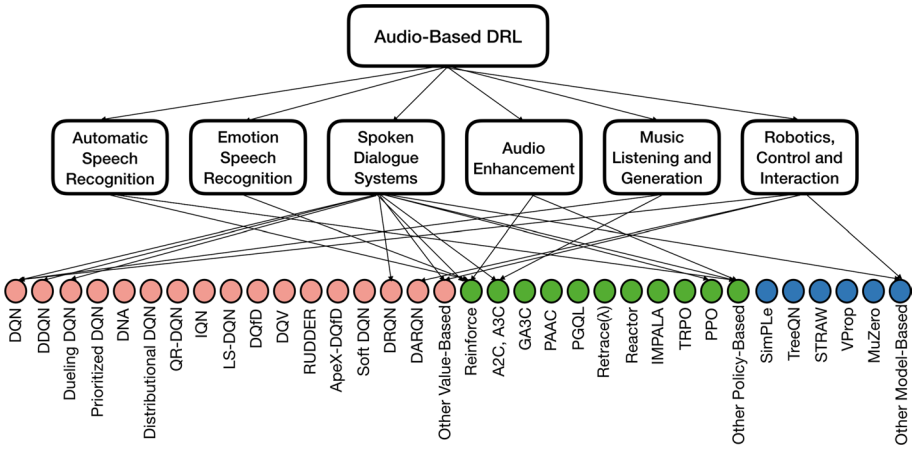

Fig. 6 A summary of audio-based DRL connecting the application areas and algorithms described in the previous two sections—the coloured circles correspond to the three groups of algorithms (from left to right: value-based, policy-based, model-based). Since the lack of connections between areas and algorithms denote no or little attention in previous works, the large amount of disconnections suggest opportunities for exploring different algorithms or more comprehensively in order to find the best algorithm(s) for different areas and tasks within each area

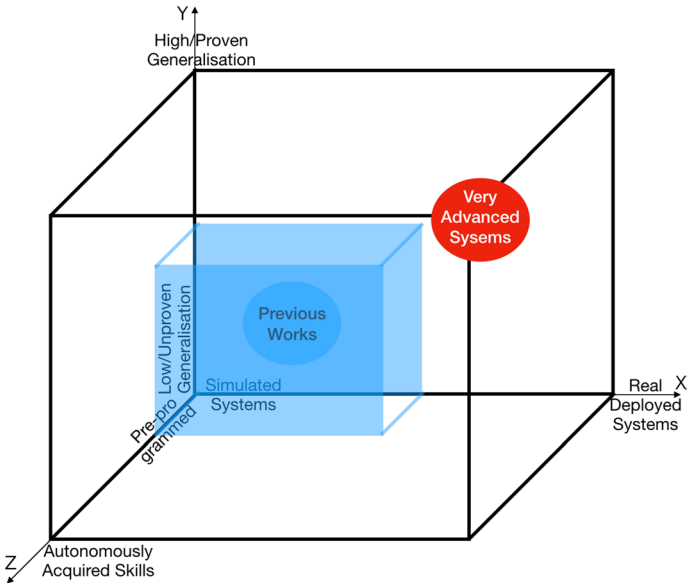

Fig. 7 A pictorial view of previous works on audio-based DRL and potential dimensions to explore in future systems. The inner cube refers to the fact that dimension Z is less developed than the other two dimensions (X,Y)

have only explored a modest space of what is possible. Based on the related works above, we have identified three main challenges that need to be addressed by future systems. Those dimensions converge in what we call 'very advanced systems'.

Table 4 Summary of research papers on dialogue systems trained with DRL algorithms

References	Application Domain(s)	DRL Algorithm	User Simulations	Transfer Learning	Training (Test) Data	Human Evaluation	Reward Function
Ammanabrolu and Riedl (2019)	Games: Slice of Life, Horror	KG-DQN	No	Yes	40 (10) games	No	+ 1 for getting closer to the finish, − 1 for extending the minimum steps, 0 otherwise
Casanueva et al. (2018)	Restaurants, laptops	FDQN, GP-Sarsa	Yes	No	4K (0.5K) dialogues	No	+20 if successful dialogue or 0 otherwise, minus dialogue length
Cuayáhuitl (2020)	Chitchat	Ensemble DQN	No	No	≤64K (1k) dialogues	Yes	+ 1 for a human-like response, − 1 for a randomly chosen response
Cuayáhuitl et al. (2019)	Robot playing games	Competitive DQN	Yes	No	20K (3K) games	Yes	+ 5 for a game win, + 1 for a draw, − 5 for a loss
Cuayáhuitl et al. (2017)	Restaurants, hotels	NDQN	Yes	No	8.7K (1K) dialogues	No	Pr(TaskSuccess) + Pr(DataLike) − number of turns × − 0.1
(Das et al. 2017)	Visual Question-Answering	Reinforce	No	No	68K (9.5K) images	Yes	Euclidean distances between predicted and target descriptions of the last 2 time steps
Fatemi et al. (2016)	Restaurants	DA2C	Yes	No	15K (0.5K) dialogues	No	+ 1 if successful dialogue, − 0.03 at each turn, − 1 if unsuccessful dialogue or hangup
Gordon-Hall et al. (2020)	MultiWoz	NDfQ	Yes	No	11.4K (1K) dialogues	No	+ 100 for successfully completing the task, − 1 at each turn

Table 4 (continued)

References	Application Domain(s)	DRL Algorithm	User Simulations	Transfer Learning	Training (Test) Data	Human Evaluation	Reward Function
Li et al. (2016)	OpenSubtitles	Reinforce	No	No	Ĩ0M (1K) sentences	Yes	Weighted scores combining ease of answering, information flow, and semantic coherence
(Lipton et al. 2018)	Movie booking	BBQN	Yes	No	20K (10K) dialogues	Yes	+ 40 if successful dialogue, − 1 at each turn, − 10 for a failed dialogue
Narasimhan et al. (2018)	Freeway, Bomberman,	Text-DQN,	No	Yes	10–15M (50K) steps	No	Learnt rewards (CNN network trained from
	Bourderchase, F &E	Text-VI					crowdsourced text descriptions of gameplays)
Peng et al. (2017)	Flights and hotels	Hierarchical DQN	Yes	No	20K (2K) dialogues	Yes	+ 120 if successful dialogue, − 1 at each turn, − 60 for a failed dialogue
Peng et al. (2018)	Movie-ticket booking	Adversarial	Yes	No	100K (5K) dialogues	No	Learnt rewards (MLP network comparing
		A2C					state-action pairs with human dialogues)
Saleh et al. (2020)	Chitchat	Hierarchical	No	No	109K (10K) dialogues	Yes	Predefined scores combining question,
		Reinforce					repetition, semantic similarity, and toxicity
Serban et al. (2017)	Chitchat	Reinforce	No	No	~5K (0.1) dialogues	Yes	Learnt rewards (linear regressor predicting
							user scores at the end of the dialogue)

Table 4 (continued)

References	Application Domain(s)	DRL Algorithm	User Simulations	Transfer Learning	Training (Test) Data	Human Evaluation	Reward Function
Su et al. (2017)	Restaurants	TRACER,	Yes	No	≤3.5K (0.6K) dialogues	No	+20 if successful dialogue (0 otherwise)
		eNACER					minus 0.05 × number of dialogue turns
Ultes et al. (2017)	Buses, restaurants,	GP-Sarsa	Yes	No	1K (0.1K) dialogues	No	Learnt rewards (Support Vector Machine
	hotels, laptops						predicting user dialogue ratings)
Xu et al. (2019)	Medical diagnosis	KR-DQN	Yes	No	423 (104) dialogues	Yes	+44 for successful diagnoses, − 22 for failed
	(4 diseases)						diagnoses, − 1 for failing symptom requests
Weisz et al. (2018)	Restaurants	ACER	Yes	No	4K (4K) dialogues	Yes	+ 20 for a successful dialogue minus number of turns in the dialogue
Williams and Zweig (2016)	Dialling domain	Reinforce	Yes	No	5K (0.5K) dialogues	No	+ 1 for successfully completing the dialogue, 0 otherwise
Zhao and Eskénazi (2016)	20-question game	DRQN	Yes	No	120K (5K) sentences	No	+ 30 for a game win, − 30 for a lost game, − 5 for a wrong guess
Zhao et al. (2019)	DealOrNotDeal,	Reinforce	Yes;No	No	≤8.4K (≤1K) dialogues	No	+≤10 for a negotiation, 0 for no agreement;
	MultiWoz						language constrained reward curve

Table 4 (continued)

References	Application Domain(s)	DRL Algorithm	User Simulations	Transfer Learning	Training (Test) Data	Human Evaluation	Reward Function
Zhang et al. (2018)	20 images	DRRN+	Yes	No	20K (1K) games	No	+10 for a game win, − 10 for a lost game,
	guessing game	DQN					a pseudo reward for question selection
Takanobu et al. (2019)	MultiWoz	GP_{mbcm}, PPO	Yes	No	10.5K (1K) dialogues	Yes	Learnt rewards (MLP network comparing
		ACER, ALDM					state-action pairs with human dialogues)
Zorrilla et al. (2021)	Restaurants	Reinforce,	Yes	No	1.6K (1.6K) dialogues	No	100*ES-50λ, if end of dialogue; 50*ES-25λ, if λ has changed; and
		Actor-Critic					− 0.1 otherwise. ES is a weighted avg. from 3 metrics and λ=1-ES.
Zhang et al. (2021)	Movie Ticket, Frames	DQN, DDQ DRD3Q, ES-DDQ	Yes	No	1.5K (128) dialogues	Yes	Learnt rewards (RNN-MLP network for predicting emotions from user responses)

Due to different experimental settings, these studies are not comparable

5.1 Real-world audio-based systems

Most of the DRL algorithms described in Sect. 3 carry out experiments on the Atari benchmark (Bellemare et al. 2013), where there is no difference between training and test environments. This is an important limitation in the literature, and it should be taken into account in the development of future DRL algorithms. Nonetheless, those efforts have been worth making progress in core DRL research, which have the potential of influencing a large amount of applications. In contrast, audio-based DRL applications tend to make use of a more explicit separation between training and test environments. While audio-based DRL agents may be trained from offline interactions or simulations, their performance requires to be assessed using a separate set of offline data or real interactions. The latter (often referred to as *human evaluations*) is very important for analysing and evidencing the quality of learnt behaviours. Learning behaviour offline is typically preferred for two main reasons: (i) large training times for inducing the best possible behaviour; and (ii) to avoid nonsensical or incoherent behaviour, due to exploration strategies used during training, unless one has is a mechanism in place to assure reasonable behaviour during online learning. In almost all (if not all) audio-based systems, the creation of data is difficult and expensive. This highlights the need for more data-efficient algorithms—especially if DRL agents are expected to learn from real data instead of synthetic data. In high-frequency audio-based control tasks, DRL agents have the requirements of learning fast and avoiding repeating the same mistake. Real-world audio-based systems require algorithms that are sample efficient and performant in their operations. This makes the application of DRL algorithms in real systems very challenging. Some studies such as (Finn et al. 2017; Chua et al. 2018; Buckman et al. 2018), have presented approaches to improve the sample efficiency of DRL systems. These approaches, however, have not been applied to audio-based systems. This suggests that much more research is required to make DRL more practical and successful for its application in real audio-based systems.

5.2 Knowledge transfer and generalisation

Learning behaviours from complex signals like speech and audio with DRL requires processing high-dimensional inputs and performing extensive training on a large number of samples to achieve improved performance. The unavailability of large labelled datasets is indeed one of the major obstacles in the area of audio-driven DRL (Purwins et al. 2019; Latif et al. 2020). Moreover, it is computationally expensive to train a single DRL agent, and there is a need for training multiple DRL agents in order to equip audio-based systems with a variety of learnt skills. Therefore, some researchers have turned their attention to studying different schemes such as policy distillation (Rusu et al. 2015), progressive neural networks (Rusu et al. 2016), multi-domain/multi-task learning (Cuayáhuitl et al. 2017; Ultes et al. 2017; Li et al. 2015; Jaderberg et al. 2016) and others (Yin and Pan 2017; Nguyen et al. 2020; Glatt et al. 2016) to promote transfer learning and generalisation in DRL to improve system performance and reduce computational costs. Only a few studies in dialogue systems have started to explore transfer learning in DRL for the speech, audio and dialogue domains (Mo et al. 2018; Carrara et al. 2018; Chen et al. 2018; Narasimhan et al. 2018; Ammanabrolu and Riedl 2019), and more research is needed in this area. When large amounts of data exist, one could opt for ignoring knowledge transfer—but most of the time this is not the case. In the presence of small or medium-size datasets, it is worth

considering the idea of transferring knowledge induced from other datasets to the one at hand. DRL agents are often trained from scratch instead of inheriting useful behaviours from other agents. Some agents from Table 4 [such as (Williams and Zweig 2016; Liu et al. 2017; Zorrilla et al. 2021)] have avoided learning from scratch by showing that applying DRL on top of non-DRL or supervised methods yields improved performance due to the optimisation element that DRL brings instead of only mimicking demonstration data. But those systems typically focus a single dataset and the idea of transferring useful and effective knowledge from other/many tasks to a new or targeted task remains to be demonstrated. Research efforts in these directions would contribute towards more practical, cost-effective, and robust applications of audio-based DRL agents. On the one hand, to train agents less data-intensively, and on the other to achieve reasonable performance in the real world.

5.3 Multi-agent and truly autonomous systems

Audio-based DRL has achieved impressive performance in single-agent domains, where the environment stays mostly stationary. But in the case of audio-based systems operating in real-world scenarios, the environments are typically challenging and dynamic. For instance, multi-lingual ASR and spoken dialogue systems need to learn policies for different languages and domains. These tasks not only involve a high degree of uncertainty and complicated dynamics but are also characterised by the fact that they are situated in the real physical world, thus have an inherently distributed nature. The problem, thus, falls naturally into the realm of multi-agent RL (MARL), an area of knowledge with a relatively long history, and has recently re-emerged due to advances in single-agent RL techniques (Littman 1994; Hernandez-Leal et al. 2019). Coupled with recent advances in DNNs, MARL has been in the limelight for many recent breakthroughs in various domains including control systems, communication networks, economics, etc. However, applications in the audio processing domain are relatively limited due to various challenges. The learning goals in MARL are multidimensional—because the objectives of all agents are not necessarily aligned. This situation can arise for example in simultaneous emotion and speaker voice recognition, where the goal of one agent is to identify emotions and the goal of the other agent is to recognise the speaker. As a consequence, these agents can independently perceive the environment, and act according to their individual objectives (rewards) thus modifying the environment. This can bring up the challenge of dealing with equilibrium points, as well as some additional performance criteria beyond return-optimisation, such as the robustness against potential adversarial agents. As all agents try to improve their policies according to their interests concurrently, therefore the action executed by one agent affects the goals and objectives of the other agents (e.g. speaker, gender, and emotion identification from the speech at the same time), and vice-versa.

One remaining challenging aspect is that of autonomous skill acquisition. Most, if not all, DRL agents currently require a substantial amount of pre-programming as opposed to acquiring skills autonomously to enable personalised/extensible behaviour. Such pre-programming includes explicit implementations of states, actions, rewards, and policies. Examples of pre-programing agents are as follows: implementing a particular combination of features derived from audio/word/sentence embeddings among others; implementing a particular set of dialogue actions instead of learned ones; implementing a particular reward function focused on optimising task success and dialogue length instead of other factors; and implementing a policy using purely learnt behaviour instead of rule-based

and DRL-based or supervised-based and DRL-based, among others. Pre-programming is needed due to not or partially knowing what the best representations are for different tasks. As agents become more advanced, those representations of states, actions, rewards and policies will be better known across tasks and therefore the amount of pre-programming will be reduced. Although substantial progress in different areas has been made, the idea of creating audio-driven DRL agents that autonomously learn their states, actions, and rewards in order to induce useful skills remains to be investigated further across applications. Such kind of agents would have to know when and how to observe their environments, identify a task and input features, induce a set of actions, induce a reward function (from audio, images, or both among others), and use all of that to train policies. Such agents have the potential to show advanced levels of intelligence, and they would be very useful for applications such as personal assistants or interactive robots.

6 Summary of audio-based DRL research and future directions

This literature review shows that DRL is becoming popular in audio processing and related applications. We collected DRL research papers in six different but related areas: automatic speech recognition (ASR), speech emotion recognition (SER), spoken dialogue systems (SDSs), audio enhancement, audio-driven robotic control, and music generation. A summary of our findings for each area is given below.

1. In ASR, most of the studies have used policy gradient-based DRL, as it allows learning an optimal policy that maximises the performance objective. We found studies aiming to solve the complexity of ASR models (Dudziak et al. 2019), tackle slow convergence issues (Williams 1992), and speed up the convergence in DRL (Rajapakshe et al. 2020).
2. The development of SDSs with DRL is gaining interest and different studies have shown very interesting results that have outperformed current state-of-the-art DL approaches (Weisz et al. 2018). However, there is still room for improvement regarding the effective and practical training of DRL-based spoken dialogue systems.
3. Several studies have also applied DRL to emotion recognition and empirically showed that DRL can (i) lower latency while making predictions (Lakomkin et al. 2018), (ii) understand emotional dynamics in communication (Sangeetha and Jayasankar 2019), and (iii) enhance human-computer interaction (Chen et al. 2017).
4. In the case of audio enhancement, studies have shown the potential of DRL. While these studies have focused their attention on the speech signals, DRL can be used to optimise the audio enhancement module along with performance objectives such as those in ASR (Shen et al. 2019).
5. In music generation, DRL can optimise rules of music theory as validated in different studies (Jaques et al. 2016; Guimaraes et al. 2017). It can also be used to search for new

tone synthesis parameters (Lan et al. 2019). Moreover, DRL can be used to perform score following to track a musical performance (Dorfer et al. 2018), and it is even suitable for tracking real piano recordings (Henkel et al. 2019), among other possible tasks.

6. In robotics, audio-based DRL agents are in their infancy. Previous studies have trained DRL-based agents using simulations, which have shown that reinforcement principles help agents in the acquisition of spoken language. Some recent works (Hussain et al. 2019, 2019) have shown that DRL can be utilised to train gaze controllers and speech-driven backchannels like laughs in human–robot interaction—and this is only the beginning of larger-scale embodied DRL-based agents.

The related works reviewed above highlight several benefits of using DRL for audio processing and applications. Challenges remain before such advancements will succeed in the real world, including endowing agents with commonsense knowledge, knowledge transfer, generalisation, and autonomous learning, among others—see Fig. 7. Such advances need to be demonstrated not only in simulated and stationary environments but in real and non-stationary ones as in real-world scenarios. Steady progress, however, is being made in the right direction for designing more adaptive audio-based systems that can be better suited for real-world settings. If such scientific progress keeps growing rapidly, perhaps we are not too far away from AI-based autonomous systems that can listen, process, and understand audio and act in more human-like ways in increasingly complex environments. In Table 5, we compare different DRL toolkits in terms of implemented algorithms, which aim to help researchers to select suitable tools to study DRL techniques.

7 Conclusions

In this work, we have focused on presenting a comprehensive review of deep reinforcement learning (DRL) techniques for audio based applications. We reviewed DRL research works in six different audio-related areas including automatic speech recognition (ASR), speech emotion recognition (SER), spoken dialogue systems (SDSs), audio enhancement, audio-driven robotic control, and music generation. In all of these areas, the use of DRL techniques is becoming increasingly popular, and ongoing research on this topic has explored many DRL algorithms with encouraging results for audio-related applications. Apart from providing a detailed review, we have also highlighted (i) various challenges that hinder DRL research in audio applications and (ii) various avenues for exciting future research. We hope that this paper will help researchers and practitioners interested in exploring and solving problems in the audio and related areas using DRL techniques.

Table 5 Comparing DRL libraries based on the state-of-the-art implemented algorithms

Tools	Available DRL algorithms										
	DQN	Cat. DQN	DDPG	NAF	SARSA	PG	AC	SAC	PPO	TD3	Reinforce
KerasRL (Plappert 2016)	✓	✗	✓	✓	✓	✗	✗	✗	✓	✗	✗
Tensorforce (Kuhnle et al. 2017)	✓	✗	✓	✓	✗	✓	✓	✗	✗	✗	✗
RL_Coach (Caspi et al. 2017)	✓	✓	✓	✗	✗	✓	✓	✓	✗	✗	✗
TF-Agents (Guadarrama et al. 2018)	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓
TF-Keras (Team 2021)	✓	✗	✓	✗	✗	✗	✓	✗	✓	✗	✗
Stable Baselines (Raffin et al. 2019)	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗
MushroomRL (D'Eramo et al. 2021)	✓	✗	✓	✓	✗	✗	✗	✓	✓	✓	✗

Data accessed on 8-August-2021

Acknowledgements We would like to thank Kai Arulkumaran (Imperial College London, United Kingdom) and Dr Soujanya Poria (Singapore University of Technology and Design) for proving feedback on the paper. We also thank Waleed Iqbal (Queen Mary University of London, United Kingdom) for helping with the extraction of DRL related data from Scopus (Fig. 1).

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

References

Abbeel P, Ng AY (2004) Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the twenty-first international conference on Machine learning, p 1

Abdel-Hamid O, Mohamed Ar, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 22(10)

Alamdari N, Lobarinas E, Kehtarnavaz N (2020) Personalization of hearing aid compression by human-in-the-loop deep reinforcement learning. IEEE Access 8:203503–203515. https://doi.org/10.1109/ACCESS.2020.3035728

Alfredo C, Humberto C, Arjun C (2017) Efficient parallel methods for deep reinforcement learning. In: The Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM)

Ali HS, ul Hassan F, Latif S, Manzoor HU, Qadir J (2021) Privacy enhanced speech emotion communication using deep learning aided edge computing. In: 2021 IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1–5. IEEE

Allan M, Williams C (2005) Harmonising chorales by probabilistic inference. In: Advances in Neural Information Processing Systems (NIPS)

Ameixa D, Coheur L, Redol RA (2013) From subtitles to human interactions: introducing the subtle corpus. Tech. rep., Tech. rep., INESC-ID (November 2014)

Ammanabrolu P, Riedl M (2019) Transfer in deep reinforcement learning using knowledge graphs. In: Ustalov D, Somasundaran S, Jansen P, Glavas G, Riedl M, Surdeanu M, Vazirgiannis M (eds) Workshop on Graph-Based Methods for Natural Language Processing, TextGraphs@EMNLP. Association for Computational Linguistics

Arjona-Medina JA, Gillhofer M, Widrich M, Unterthiner T, Brandstetter J, Hochreiter S (2019) Rudder: Return decomposition for delayed rewards. In: Advances in Neural Information Processing Systems (NIPS)

Arora G, Rahimi A, Baldwin T (2019) Does an lstm forget more than a cnn? an empirical study of catastrophic forgetting in nlp. In: Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association, pp. 77–86

Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA (2017) Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine 34(6)

Asri LE, Schulz H, Sharma S, Zumer J, Harris J, Fine E, Mehrotra R, Suleman K (2017) Frames: a corpus for adding memory to goal-oriented dialogue systems. In: Jokinen K, Stede M, DeVault D, Louis A (eds) Annual SIGdial Meeting on Discourse and Dialogue. ACL

Babaeizadeh M, Frosio I, Tyree S, Clemons J, Kautz J (2017) Reinforcement learning through asynchronous advantage actor-critic on a gpu. In: Learning Representations. ICLR

Baby D, Gemmeke JF, Virtanen T, et al (2015) Exemplar-based speech enhancement for deep neural network based automatic speech recognition. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Bae JS, Bak TJ, Joo YSy, Cho HY (2021) Hierarchical context-aware transformers for non-autoregressive text to speech. arXiv preprint arXiv:2106.15144

Barker J, Marxer R, Vincent E, Watanabe S (2015) The third 'CHiME'speech separation and recognition challenge: Dataset, task and baselines. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)

Bellemare MG, Dabney W, Munos R (2017) A distributional perspective on reinforcement learning. In: International Conference on Machine Learning (ICML). JMLR. org

Bellemare MG, Naddaf Y, Veness J, Bowling M (2013) The Arcade learning environment: An evaluation platform for general agents. J. Artif. Intell. Res. 47

Bellman R (1966) Dynamic programming. Science 153(3731)

Buckman J, Hafner D, Tucker G, Brevdo E, Lee H (2018) Sample-efficient reinforcement learning with stochastic ensemble value expansion. In: Advances in Neural Information Processing Systems (NIPS)

Budzianowski P, Ultes S, Su P, Mrksic N, Wen T, Casanueva I, Rojas-Barahona LM, Gasic M (2017) Subdomain modelling for dialogue management with hierarchical reinforcement learning. In: K. Jokinen, M. Stede, D. DeVault, A. Louis (eds.) Annual SIGdial Meeting on Discourse and Dialogue. ACL

Budzianowski P, Wen TH, Tseng BH, Casanueva I, Ultes S, Ramadan O, Gasic M (2018) Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In: Conference on Empirical Methods in Natural Language Processing (EMNLP)

Bui H, Chong NY (2019) Autonomous speech volume control for social robots in a noisy environment using deep reinforcement learning. In: IEEE International Conference on Robotics and Biomimetics (ROBIO)

Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss, B (2005) A database of german emotional speech. In: European Conference on Speech Communication and Technology

Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation 42(4)

Busso C, Parthasarathy S, Burmania A, AbdelWahab M, Sadoughi N, Provost EM (2016) MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. IEEE Transactions on Affective Computing 8(1)

Carrara N, Laroche R, Bouraoui JL, Urvoy T, Pietquin O (2018) Safe transfer learning for dialogue applications

Carrara N, Laroche R, Pietquin O (2017) Online learning and transfer for user adaptation in dialogue systems. In: SIGDIAL/SEMDIAL joint special session on negotiation dialog 2017

Casanueva I, Budzianowski P, Su PH, Mrkšić N, Wen TH, Ultes S, Rojas-Barahona L, Young S, Gašić M (2017) A benchmarking environment for reinforcement learning based task oriented dialogue management. Deep Reinforcement Learning Symposium, NIPS

Casanueva I, Budzianowski P, Su P, Ultes S, Rojas-Barahona LM, Tseng B, Gasic M (2018) Feudal reinforcement learning for dialogue management in large domains. In: M.A. Walker, H. Ji, A. Stent (eds.) North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)

Caspi I, Leibovich G, Novik G, Endrawis S (2017). Reinforcement learning coach. https://doi.org/10.5281/zenodo.1134899

Chang SY, Li B, Simko G, Sainath TN, Tripathi A, van den Oord A, Vinyals O (2018) Temporal modeling using dilated convolution and gating for voice-activity-detection. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Chan W, Jaitly N, Le Q, Vinyals O (2016) Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Chen L, Chang C, Chen Z, Tan B, Gasic M, Yu K (2018) Policy adaptation for deep reinforcement learning-based dialogue management. In: IEEE International Conference on Acoustics, Speech and Signal ICASSP

Chen Z, Chen L, Zhou X, Yu K (2020) Deep reinforcement learning for on-line dialogue state tracking. arXiv preprint arXiv:2009.10321

Chen Y, Guo Q, Liang X, Wang J, Qian Y (2019) Environmental sound classification with dilated convolutions. Applied Acoustics 148

Chen C, Jain U, Schissler C, Gari SVA., Al-Halah Z, Ithapu VK, Robinson P, Grauman K (2019) Audio-visual embodied navigation. environment 97, 103

Chen C, Majumder S, Al-Halah Z, Gao R, Ramakrishnan SK, Grauman K (2020) Learning to set waypoints for audio-visual navigation. In: International Conference on Learning Representations

Chen M, Wang S, Liang PP, Baltrušaitis T, Zadeh A, Morency LP (2017) Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: ACM International Conference on Multimodal Interaction

Chi PH, Chung PH, Wu TH, Hsieh CC, Chen YH, Li SW, Lee Hy (2021) Audio albert: A lite bert for self-supervised learning of audio representation. In: 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 344–350. IEEE

Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP)

Chua K, Calandra R, McAllister R, Levine S (2018) Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In: Advances in Neural Information Processing Systems (NIPS)

Chung H, Jeon HB, Park JG (2020) Semi-supervised training for sequence-to-sequence speech recognition using reinforcement learning. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE

Chung H, Lee SH, Lee SW (2021) Reinforce-aligner: Reinforcement alignment search for robust end-to-end text-to-speech. arXiv preprint arXiv:2106.02830

Clark-Turner M, Begum M (2018) Deep reinforcement learning of abstract reasoning from demonstrations. In: ACM/IEEE International Conference on human–robot Interaction

Cuayáhuitl H (2009) Hierarchical reinforcement learning for spoken dialogue systes. Ph.D. thesis, University of Edinburgh

Cuayáhuitl H (2017) Simpleds: A simple deep reinforcement learning dialogue system. In: Dialogues with social robots. Springer

Cuayáhuitl H (2020) A data-efficient deep learning approach for deployable multimodal social robots. Neurocomputing 396

Cuayáhuitl H, Lee D, Ryu S, Cho Y, Choi S, Indurthi SR, Yu S, Choi H, Hwang I, Kim J (2019) Ensemble-based deep reinforcement learning for chatbots. Neurocomputing 366

Cuayáhuitl H, Renals S, Lemon O, Shimodaira H (2010) Evaluation of a hierarchical reinforcement learning spoken dialogue system. Comput. Speech Lang. 24(2)

Cuayáhuitl H, Yu S, Williamson A, Carse J (2016) Deep reinforcement learning for multi-domain dialogue systems. NIPS Workshop on Deep Reinforcement Learning

Cuayáhuitl H, Yu S, Williamson A, Carse J (2017) Scaling up deep reinforcement learning for multi-domain dialogue systems. In: International Joint Conference on Neural Networks, IJCNN

Dabney W, Ostrovski G, Silver D, Munos R (2018) Implicit quantile networks for distributional reinforcement learning. In: International Conference on Machine Learning

Dabney W, Rowland M, Bellemare MG, Munos R (2018) Distributional reinforcement learning with quantile regression. In: AAAI Conference on Artificial Intelligence

Das A, Kottur S, Moura JMF, Lee S, Batra D (2017) Learning cooperative visual dialog agents with deep reinforcement learning. In: IEEE International Conference on Computer Vision, ICCV

D'Eramo C, Tateo D, Bonarini A, Restelli M, Peters J (2021) MushroomRL: Simplifying reinforcement learning research. Journal of Machine Learning Research 22(131), 1–5 . http://jmlr.org/papers/v22/18-056.html

Dethlefs N, Cuayáhuitl H (2015) Hierarchical reinforcement learning for situated natural language generation. Nat. Lang. Eng. 21(3)

Dorfer M, Henkel F, Widmer G (2018) Learning to listen, read, and follow: Score following as a reinforcement learning game. International Society for Music Information Retrieval Conference

Duan Y, Schulman J, Chen X, Bartlett PL, Sutskever I, Abbeel P (2016) RI^2: Fast reinforcement learning via slow reinforcement learning. arXiv preprint arXiv:1611.02779

Dudziak Ł, Abdelfattah MS, Vipperla R, Laskaridis S, Lane ND (2019) ShrinkML: End-to-end asr model compression using reinforcement learning. In: Interspeech

Ebcioğlu K (1988) An expert system for harmonizing four-part chorales. Computer Music Journal 12(3)

Emiya V, Vincent E, Harlander N, Hohmann V (2011) Subjective and objective quality assessment of audio source separation. IEEE Transactions on Audio, Speech, and Language Processing 19(7)

Espeholt L, Soyer H, Munos R, Simonyan K, Mnih V, Ward T, Doron Y, Firoiu V, Harley T, Dunning I et al (2018) IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In: International Conference on Machine Learning (ICML)

Fakoor R, He X, Tashev I, Zarar S (2017) Reinforcement learning to adapt speech enhancement to instantaneous input signal quality. Machine Learning for Audio Signal Processing workshop, NIPS

Fatemi M, Asri LE, Schulz H, He J, Suleman K (2016) Policy networks with two-stage training for dialogue systems. In: Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)

Fazel-Zarandi M, Li SW, Cao J, Casale J, Henderson P, Whitney D, Geramifard A (2017) Learning robust dialog policies in noisy environments. Workshop on Conversational AI, NIPS

Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning (ICML)

Fryen T, Eppe M, Nguyen PDH., Gerkmann T, Wermter S (2020) Reinforcement learning with time-dependent goals for robotic musicians. CoRR abs/2011.05715
Gan C, Zhang Y, Wu J, Gong B, Tenenbaum JB (2020) Look, listen, and act: Towards audio-visual embodied navigation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 9701–9707. IEEE
Gao J, Galley M, Li L (2019) Neural approaches to conversational AI. Found. Trends Inf. Retr. 13(2-3)
Gao S, Hou W, Tanaka T, Shinozaki T (2020) Spoken language acquisition based on reinforcement learning and word unit segmentation. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)
Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett, DS (1993) DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon technical report n 93
Gašić M, Young S (2013) Gaussian processes for POMDP-based dialogue manager optimization. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22(1)
Gebru ID, Ba S, Li X, Horaud R (2017) Audio-visual speaker diarization based on spatiotemporal Bayesian fusion. IEEE transactions on pattern analysis and machine intelligence 40(5)
Ghosal D, Kolekar MH (2018) Music genre recognition using deep neural networks and transfer learning. In: Interspeech, vol. 2018
Giannakopoulos P, Pikrakis A, Cotronis Y (2021) A deep reinforcement learning approach to audio-based navigation in a multi-speaker environment. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3475–3479. IEEE
Glatt R, Da Silva FL, Costa AHR (2016) Towards knowledge transfer in deep reinforcement learning. In: Brazilian Conference on Intelligent Systems (BRACIS)
Godfrey JJ, Holliman EC, McDaniel J (1992) SWITCHBOARD: Telephone speech corpus for research and development. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1
Gonzalez-Billandon J, Grasse L, Tata M, Sciutti A, Rea F (2020) Self-supervised reinforcement learning for speaker localisation with the icub humanoid robot. arXiv preprint arXiv:2011.06544
Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS)
Goodrich MA, Schultz AC (2007) human–robot interaction: a survey. Foundations and trends in human-computer interaction 1(3)
Gordon-Hall G, Gorinski PJ, Cohen SB (2020) Learning dialog policies from weak demonstrations. In: Jurafsky D, Chai J, Schluter N, Tetreault JR (eds) Annual Meeting of the Association for Computational Linguistics ACL. ACL
Graves A (2012) Sequence transduction with recurrent neural networks. Workshop on Representation Learning, International Conference of Machine Learning (ICML) 2012
Gruslys A, Azar MG, Bellemare MG, Munos R (2017) The reactor: A sample-efficient actor-critic architecture. arXiv preprint arXiv:1704.04651
Guadarrama S, Korattikara A, Ramirez O, Castro P, Holly E, Fishman S, Wang K, Gonina E, Wu N, Kokiopoulou E, Sbaiz L, Smith J, Bartók G, Berent J, Harris C, Vanhoucke V, Brevdo E (2018) TF-Agents: A library for reinforcement learning in tensorflow. https://github.com/tensorflow/agents . [Online; accessed 25-June-2019]
Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Farias PLC, Aspuru-Guzik A (2017) Objective-reinforced generative adversarial networks (organ) for sequence generation models. arXiv preprint arXiv:1705.10843
Hausknecht M, Stone P (2015) Deep recurrent Q-learning for partially observable MDPs. In: AAAI Fall Symposium Series
Haydari A, Yilmaz Y (2020) Deep reinforcement learning for intelligent transportation systems: A survey. arXiv preprint arXiv:2005.00935
He Y, Lin J, Liu Z, Wang H, Li LJ, Han S: Amc: Automl for model compression and acceleration on mobile devices. In: European Conference on Computer Vision (ECCV) (2018)
Henderson M, Thomson B, Williams JD (2014) The third dialog state tracking challenge. In: 2014 IEEE Spoken Language Technology Workshop (SLT), pp. 324–329. IEEE
Henderson M, Thomson B, Williams JD: The second dialog state tracking challenge. In: Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL), pp. 263–272 (2014)
Henkel F, Balke S, Dorfer M, Widmer G (2019) Score following as a multi-modal reinforcement learning problem. Transactions of the International Society for Music Information Retrieval 2(1)

Hermann KM, Hill F, Green S, Wang F, Faulkner R, Soyer H, Szepesvari D, Czarnecki WM, Jaderberg M, Teplyashin D, et al (2017) Grounded language learning in a simulated 3D world. arXiv preprint arXiv:1706.06551

Hernandez-Leal P, Kartal B, Taylor ME (2019) A survey and critique of multiagent deep reinforcement learning. Autonomous Agents and Multi-Agent Systems 33(6)

Hester T, Vecerik M, Pietquin O, Lanctot M, Schaul T, Piot B, Horgan D, Quan J, Sendonaris A, Osband I et al (2018) Deep Q-learning from demonstrations. In: AAAI Conference

Heusser V, Freymuth N, Constantin S, Waibel A (2019) Bimodal speech emotion recognition using pre-trained language models. arXiv preprint arXiv:1912.02610

Hill F, Hermann KM, Blunsom P, Clark S (2018) Understanding grounded language learning agents

Hinton G, Deng L, Yu D, Dahl GE, Mohamed Ar, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN et al (2012) Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine 29(6)

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural computation 9(8)

Howard N, Cambria E (2013) Intention awareness: Improving upon situation awareness in human-centric environments. Human-centric Computing and Information Sciences 3(9)

Hsu WN, Zhang Y, Glass J (2017) Learning latent representations for speech generation and transformation. In: Interspeech

Huang KY, Wu CH, Hong QB, Su MH, Chen YH (2019) Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5866–5870. IEEE

Hussain N, Erzin E, Sezgin TM, Yemez Y (2019) Batch recurrent Q-learning for backchannel generation towards engaging agents. In: International Conference on Affective Computing and Intelligent Interaction (ACII)

Hussain N, Erzin E, Sezgin TM, Yemez Y (2019) Speech driven backchannel generation using deep q-network for enhancing engagement in human–robot interaction. In: Interspeech

Jaderberg M, Mnih V, Czarnecki WM, Schaul T, Leibo JZ, Silver D, Kavukcuoglu K (2016) Reinforcement learning with unsupervised auxiliary tasks. International Conference on Learning Representations (ICLR)

Jaitly N, Le QV, Vinyals O, Sutskever I, Sussillo D, Bengio S (2016) An online sequence-to-sequence model using partial conditioning. In: Advances in Neural Information Processing Systems (NIPS)

Jaques N, Gu S, Turner RE, Eck D (2016) Generating music by fine-tuning recurrent neural networks with reinforcement learning

Jiang N, Jin S, Duan Z, Zhang C (2020) Rl-duet: Online music accompaniment generation using deep reinforcement learning. Proceedings of the AAAI Conference on Artificial Intelligence 34:710–718

Kaiser Ł, Babaeizadeh M, Miłos P, Osiński B, Campbell RH, Czechowski K, Erhan D, Finn C, Kozakowski P, Levine S et al (2019) Model based reinforcement learning for atari. In: International Conference on Learning Representations

Kala T, Shinozaki T (2018) Reinforcement learning of speech recognition system based on policy gradient and hypothesis selection. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Karita S, Chen N, Hayashi T, Hori T, Inaguma H, Jiang Z, Someki M, Soplin NEY, Yamamoto R, Wang X et al (2019) A comparative study on transformer vs rnn in speech applications. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 449–456. IEEE

Karita S, Ogawa A, Delcroix M, Nakatani T (2018) Sequence training of encoder-decoder model using policy gradient for end-to-end speech recognition. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Kingma DP, Welling M (2013) Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114

Kiran BR, Sobh I, Talpaert V, Mannion P, Sallab AAA, Yogamani S, Pérez P (2020) Deep reinforcement learning for autonomous driving: A survey. arXiv preprint arXiv:2002.00444

Kohl N, Stone P (2004) Policy gradient reinforcement learning for fast quadrupedal locomotion. In: IEEE International Conference on Robotics and Automation (ICRA), vol. 3

Koizumi Y, Niwa K, Hioka Y, Kobayashi K, Haneda Y (2017) DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Konda VR, Tsitsiklis JN (1999) Actor-Critic agorithms. In: Neural Information Processing Systems (NIPS)

Kotecha N (2018) Bach2Bach: Generating music using a deep reinforcement learning approach. arXiv preprint arXiv:1812.01060

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS)

Krueger B (2016) Classical piano midi page
Kuhnle A, Schaarschmidt M, Fricke K (2017) Tensorforce: a tensorflow library for applied reinforcement learning. Web page . https://github.com/tensorforce/tensorforce
Lakomkin E, Zamani MA, Weber C, Magg S, Wermter S (2018) Emorl: continuous acoustic emotion classification using deep reinforcement learning. In: IEEE International Conference on Robotics and Automation (ICRA)
Łańcucki A (2021) Fastpitch: Parallel text-to-speech with pitch prediction. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6588–6592. IEEE
Lange S, Riedmiller MA, Voigtländer A (2012) Autonomous reinforcement learning on raw visual input data in a real world application. In: International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, June 10-15, 2012. IEEE
Lan Q, Tørresen J, Jensenius AR (2019) RaveForce: A deep reinforcement learning environment for music. In: Proc. of the SMC Conferences. Society for Sound and Music Computing
Lathuilière S, Massé B, Mesejo P, Horaud R (2018) Deep reinforcement learning for audio-visual gaze control. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Lathuilière S, Massé B, Mesejo P, Horaud R (2019) Neural network based reinforcement learning for audio–visual gaze control in human–robot interaction. Pattern Recognition Letters 118
Latif S (2020) Deep representation learning for improving speech emotion recognition
Latif S, Rana R, Khalifa S, Jurdak R, Schuller BW (2020) Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition. Proc. Interspeech 2020:2327–2331
Latif S, Asim M, Rana R, Khalifa S, Jurdak R, Schuller BW (2020) Augmenting generative adversarial networks for speech emotion recognition. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2020:521–525
Latif S, Kim I, Calapodescu I, Besacier L (2021) Controlling prosody in end-to-end tts: A case study on contrastive focus generation. In: Proceedings of the 25th Conference on Computational Natural Language Learning, pp. 544–551
Latif S, Qadir J, Bilal M (2019) Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition. In: International Conference on Affective Computing and Intelligent Interaction (ACII)
Latif S, Qadir J, Qayyum A, Usama M, Younis S (2020) Speech technology for healthcare: Opportunities, challenges, and state of the art. IEEE Reviews in Biomedical Engineering
Latif S, Rana R, Khalifa S, Jurdak R, Epps J (2019) Direct modelling of speech emotion from raw speech. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association INTERSPEECH 2019, pp. 3920–3924. International Speech Communication Association
Latif S, Rana R, Khalifa S, Jurdak R, Epps J, Schuller BW (2020) Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. IEEE Transactions on Affective Computing
Latif S, Rana R, Khalifa S, Jurdak R, Qadir J, Schuller BW (2020) Deep representation learning in speech processing: Challenges, recent advances, and future trends. arXiv preprint arXiv:2001.00378
Latif S, Rana R, Khalifa S, Jurdak R, Qadir J, Schuller BW (2021) Survey of deep representation learning for speech emotion recognition. IEEE Transactions on Affective Computing
Latif S, Rana R, Khalifa S, Jurdak R, Schuller BW (2022) Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition. IEEE Transactions on Affective Computing
Latif S, Rana R, Qadir J (2018) Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness. arXiv preprint arXiv:1811.11402
Latif S, Rana R, Qadir J, Epps J (2018) Variational autoencoders for learning latent representations of speech emotion: A preliminary study. In: Interspeech
Lawson D, Chiu CC, Tucker G, Raffel C, Swersky K, Jaitly N (2018) Learning hard alignments with variational inference. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)
LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural computation 1(4)
Lee Sg, Hwang U, Min S, Yoon S (2017) Polyphonic music generation with sequence generative adversarial networks. arXiv preprint arXiv:1710.11418
Le N, Rathour VS, Yamazaki K, Luu K, Savvides M (2021) Deep reinforcement learning in computer vision: a comprehensive survey. Artificial Intelligence Review pp. 1–87
Levine S, Finn C, Darrell T, Abbeel P (2016) End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research 17(1)

Levine N, Zahavy T, Mankowitz DJ, Tamar A, Mannor S (2017) Shallow updates for deep reinforcement learning. In: Advances in Neural Information Processing Systems (NIPS)
Levin E, Pieraccini R, Eckert W (2000) A stochastic model of human-machine interaction for learning dialog strategies. IEEE Transactions Speech Audio Process. 8(1)
Li Y (2017) Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274
Li J, Deng L, Haeb-Umbach R, Gong Y (2015) Robust automatic speech recognition: a bridge to practical applications. Academic Press
Li X, Li L, Gao J, He X, Chen J, Deng L, He J (2015) Recurrent reinforcement learning: a hybrid approach. arXiv preprint arXiv:1509.03044
Li J, Mohamed A, Zweig G, Gong Y (2015) LSTM time and frequency recurrence for automatic speech recognition. In: IEEE workshop on automatic speech recognition and understanding (ASRU)
Li J, Monroe W, Ritter A, Galley M, Gao J, Jurafsky D (2016) Deep reinforcement learning for dialogue generation. CoRR abs/1606.01541
Lin T, Wang Y, Liu X, Qiu X (2021) A survey of transformers. arXiv preprint arXiv:2106.04554
Lipton ZC (2015) A critical review of recurrent neural networks for sequence learning. CoRR abs/1506.00019
Lipton ZC, Li X, Gao J, Li L, Ahmed F, Deng L (2018) Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In: S.A. McIlraith KQ Weinberger (eds.) AAAI Conference on Artificial Intelligence
Li B, Tsao Y, Sim KC (2013) An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition. In: Interspeech
Littman ML (1994) Markov games as a framework for multi-agent reinforcement learning. In: Machine learning proceedings 1994. Elsevier
Liu R, Sisman B, Li H (2021) Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability. arXiv preprint arXiv:2104.01408
Liu B, Tur G, Hakkani-Tur D, Shah P, Heck L (2017) End-to-end optimization of task-oriented dialogue model with deep reinforcement learning. In: NIPS Workshop on Conversational AI
Liu R, Yang J, Liu M (2019) A new end-to-end long-time speech synthesis system based on tacotron2. In: International Symposium on Signal Processing Systems
Luo Y, Chiu CC, Jaitly N, Sutskever I: Learning online alignments with continuous rewards policy gradient. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017)
Luong NC, Hoang DT, Gong S, Niyato D, Wang P, Liang, YC, Kim DI (2019) Applications of deep reinforcement learning in communications and networking: A survey. IEEE Communications Surveys & Tutorials 21(4)
Lu L, Zhang X, Renals S (2016) On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)
Maciejewski M, Wichern G, McQuinn E, Le Roux J (2020) WHAMR!: Noisy and reverberant single-channel speech separation. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)
Majumder S, Al-Halah Z, Grauman K(2021) Move2hear: Active audio-visual source separation. arXiv preprint arXiv:2105.07142
Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E (2019) DialogueRNN: An attentive RNN for emotion detection in conversations. In: AAAI Conference on Artificial Intelligence, vol. 33
Ma S, McDuff D, Song Y (2019) M3D-GAN: Multi-modal multi-domain translation with universal attention. arXiv preprint arXiv:1907.04378
Mamun N, Khorram S, Hansen JH (2019) Convolutional neural network-based speech enhancement for cochlear implant recipients. In: Interspeech
Ma Y, Nguyen KL, Xing F, Cambria E (2020) A survey on empathetic dialogue systems. Information Fusion 64
McKeown G, Valstar M, Cowie R, Pantic M, Schroder M (2011) The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE transactions on affective computing 3(1)
Misra DK, Sung J, Lee K, Saxena A (2016) Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. Int. J. Robotics Res. 35(1-3)
Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K (2016) Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning (ICML)

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540)
Mohamed Ar, Dahl G, Hinton G (2009) Deep belief networks for phone recognition. In: NIPS workshop on deep learning for speech recognition and related applications
Mohan DSR, Lenain R, Foglianti L, Teh TH, Staib M, Torresquintero A, Gao J (2020) Incremental text to speech for neural sequence-to-sequence models using reinforcement learning. Proc. Interspeech 2020:3186–3190
Moreira I, Rivas J, Cruz F, Dazeley R, Ayala A, Fernandes BJT (2020) Deep reinforcement learning with interactive feedback in a human–robot environment. CoRR abs/2007.03363
Mo K, Zhang Y, Li S, Li J, Yang Q (2018) Personalizing a dialogue system with transfer reinforcement learning. In: AAAI Conference
Munos R, Stepleton T, Harutyunyan A, Bellemare M (2016) Safe and efficient off-policy reinforcement learning. In: Advances in Neural Information Processing Systems (NIPS)
Naeem M, Rizvi STH, Coronato A (2020) A gentle introduction to reinforcement learning and its application in different fields. IEEE Access 8:209320–209344
Narasimhan K, Barzilay R, Jaakkola TS (2018) Grounding language for transfer in deep reinforcement learning. J. Artif. Intell. Res. 63
Nardelli N, Synnaeve G, Lin Z, Kohli P, Torr PH, Usunier, N (2018) Value propagation networks. In: International Conference on Learning Representations
Ng AY, Coates A, Diel M, Ganapathi V, Schulte J, Tse B, Berger E, Liang E (2006) Autonomous inverted helicopter flight via reinforcement learning. In: Experimental robotics IX. Springer
Ng AY, Russell SJ, et al (2000) Algorithms for inverse reinforcement learning. In: Icml, vol. 1, p. 2
Nguyen ND, Nguyen T, Nahavandi S (2017) System design perspective for human-level agents using deep reinforcement learning: A survey. IEEE Access 5
Nguyen TT, Nguyen ND, Nahavandi S (2020) Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. IEEE transactions on cybernetics
Ntalampiras S (2017) A transfer learning framework for predicting the emotional content of generalized sound events. The Journal of the Acoustical Society of America 141(3):1694–1701
Ntalampiras S (2018) Bird species identification via transfer learning from music genres. Eco Inform 44:76–81
Ntalampiras S (2021) Speech emotion recognition via learning analogies. Pattern Recogn Lett 144:21–26
O'Donoghue B, Munos R, Kavukcuoglu K, Mnih V (2016) PGQ: Combining policy gradient and Q-learning. arXiv preprint arXiv:1611.01626
Oh J, Chockalingam V, Lee H et al (2016) Control of memory, active perception, and action in minecraft. In: International Conference on Machine Learning
Oh J, Singh S, Lee H (2017) Value prediction network. In: Advances in Neural Information Processing Systems (NIPS)
Ouyang X, Nagisetty S, Goh EGH, Shen S, Ding W, Ming H, Huang DY (2018) Audio-visual emotion recognition with capsule-like feature representation and model-based reinforcement learning. In: 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), pp. 1–6. IEEE
Paek T (2006) Reinforcement learning for spoken dialogue systems: Comparing strengths and weaknesses for practical deployment. In: Proc. Dialog-on-Dialog Workshop, Interspeech. Citeseer
Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: an asr corpus based on public domain audio books. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)
Parisotto E, Salakhutdinov R (2018) Neural map: Structured memory for deep reinforcement learning. In: International Conference on Learning Representations
Paul DB, Baker JM (1992) The design for the wall street journal-based CSR corpus. In: Workshop on Speech and Natural Language. ACL
Peng B, Li X, Gao J, Liu J, Chen Y, Wong K (2018) Adversarial advantage actor-critic model for task-completion dialogue policy learning. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
Peng B, Li X, Li L, Gao J, Çelikyilmaz A, Lee S, Wong K (2017) Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In: M. Palmer, R. Hwa, S. Riedel (eds.) Conference on Empirical Methods in Natural Language Processing EMNLP. ACL
Pham N, Nguyen T, Niehues J, Müller M, Waibel A (2019) Very deep self-attention networks for end-to-end speech recognition. In: Kubin G, Kacic Z (eds) Interspeech. ISCA
Plappert M (2016) Keras-RL. https://github.com/keras-rl/keras-rl

Pohlen T, Piot B, Hester T, Azar MG, Horgan D, Budden D, Barth-Maron G, Van Hasselt H, Quan J, Večerík, M et al (2018) Observe and look further: Achieving consistent performance on Atari. arXiv preprint arXiv:1805.11593

Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R (2019) MELD: A multimodal multi-party dataset for emotion recognition in conversations. In: Annual Meeting of the Association for Computational Linguistics ACL

Poria S, Majumder N, Mihalcea R, Hovy E (2019) Emotion recognition in conversation: Research challenges, datasets, and recent advances. IEEE Access 7

Purwins H, Li B, Virtanen T, Schlüter J, Chang SY (2019) Sainath T, Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing 13(2)

Qian Y, Bi M, Tan T, Yu K (2016) Very deep convolutional neural networks for noise robust speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24(12)

Qureshi AH, Nakamura Y, Yoshikawa Y, Ishiguro H (2018) Intrinsically motivated reinforcement learning for human–robot interaction in the real-world. Neural Networks 107

Radzikowski K, Nowak R, Wang L, Yoshie O: Dual supervised learning for non-native speech recognition. EURASIP Journal on Audio, Speech, and Music Processing 2019(1) (2019)

Raffel C, Luong MT, Liu PJ, Weiss RJ, Eck D (2017) Online and linear-time attention by enforcing monotonic alignments. In: International Conference on Machine Learning (ICML). JMLR. org

Raffin A, Hill A, Ernestus M, Gleave A, Kanervisto A, Dormann N (2019) Stable baselines3. https://github.com/DLR-RM/stable-baselines3

Rajapakshe T, Latif S, Rana R, Khalifa S, Schuller BW: Deep reinforcement learning with pre-training for time-efficient training of automatic speech recognition. arXiv preprint arXiv:2005.11172 (2020)

Rastogi A, Zang X, Sunkara S, Gupta R, Khaitan P (2020)Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI. AAAI Press

Rath SP, Povey D, Veselý K, Cernocký J (2013) Improved feature processing for deep neural networks. In: Interspeech. ISCA

Ravindran B (2019) Introduction to deep reinforcement learning

Recommendation IT (2001) Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU-T p, Rec, p 862

Ren Y, Ruan Y, Tan X, Qin T, Zhao S, Zhao Z, Liu TY (2019) Fastspeech: fast, robust and controllable text to speech. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 3171–3180

Rousseau A, Deléglise P, Esteve Y (2012) TED-LIUM: an automatic speech recognition dedicated corpus. In: LREC

Rusu AA, Colmenarejo SG, Gulcehre C, Desjardins G, Kirkpatrick J, Pascanu R, Mnih V, Kavukcuoglu K, Hadsell R (2015) Policy distillation. arXiv preprint arXiv:1511.06295

Rusu AA, Rabinowitz NC, Desjardins G, Soyer H, Kirkpatrick J, Kavukcuoglu K, Pascanu R, Hadsell R (2016) Progressive neural networks. NIPS Deep Learning Symposium recommendation

Sabatelli M, Louppe G, Geurts P, Wiering M (2018) Deep quality value (dqv) learning. Advances in Neural Information Processing Systems (NIPS)

Sainath TN, Li B (2016) Modeling time-frequency patterns with lstm vs. convolutional architectures for lvcsr tasks. In: Interspeech

Saleh A, Jaques N, Ghandeharioun A, Shen JH, Picard RW (2020) Hierarchical reinforcement learning for open-domain dialog. In: AAAI Conference on Artificial Intelligence

Sallab AE, Abdou M, Perot E, Yogamani S (2017) Deep reinforcement learning framework for autonomous driving. Electronic Imaging 2017(19)

Sangeetha J, Jayasankar T (2019) Emotion speech recognition based on adaptive fractional deep belief network and reinforcement learning. In: Cognitive Informatics and Soft Computing. Springer

Scalise R, Li S, Admoni H, Rosenthal S, Srinivasa SS (2018) Natural language instructions for human–robot collaborative manipulation. Int. J. Robotics Res. 37(6)

Schatzmann J, Weilhammer K, Stuttle MN, Young SJ (2006) A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. Knowledge Eng. Review 21(2)

Schaul T, Quan J, Antonoglou I, Silver D (2016) Prioritized experience replay. International Conference on Learning Representations (ICLR)

Schlüter J, Böck S (2014) Improved musical onset detection with convolutional neural networks. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Schrittwieser J, Antonoglou I, Hubert T, Simonyan K, Sifre L, Schmitt S, Guez A, Lockhart E, Hassabis D, Graepel T, et al (2019) Mastering Atari, Go, Chess and Shogi by planning with a learned model. arXiv preprint arXiv:1911.08265

Schulman J, Chen X, Abbeel P (2017) Equivalence between policy gradients and soft q-learning. arXiv preprint arXiv:1704.06440

Schulman J, Levine S, Abbeel P, Jordan M, Moritz P (2015) Trust region policy optimization. In: International Conference on Machine Learning (ICML)

Schulman J, Moritz P, Levine S, Jordan M, Abbeel P (2016) High-dimensional continuous control using generalized advantage estimation. International Conference on Learning Representations (ICLR)

Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347

Serban IV, Lowe R, Henderson P, Charlin L, Pineau J (2018) A survey of available corpora for building data-driven dialogue systems: The journal version. Dialogue Discourse 9(1)

Serban IV, Sankar C, Germain M, Zhang S, Lin Z, Subramanian S, Kim T, Pieper M, Chandar S, Ke NR, et al (2017) A deep reinforcement learning chatbot. arXiv preprint arXiv:1709.02349

Serban IV, Sankar C, Germain M, Zhang S, Lin Z, Subramanian S, Kim T, Pieper M, Chandar S, Ke NR, Mudumba S, de Brébisson A, Sotelo J, Suhubdy D, Michalski V, Nguyen A, Pineau J, Bengio Y (2017) A deep reinforcement learning chatbot. CoRR abs/1709.02349

Seurin M, Strub F, Preux P, Pietquin O (2020) A machine of few words interactive speaker recognition with reinforcement learning. In: Conference of the International Speech Communication Association (INTERSPEECH)

Shah P, Fiser M, Faust A, Kew JC, Hakkani-Tur D (2018) Follownet: Robot navigation by following natural language directions with deep reinforcement learning. arXiv preprint arXiv:1805.06150

Shannon M, Zen H, Byrne W (2012) Autoregressive models for statistical parametric speech synthesis. IEEE transactions on audio, speech, and language processing 21(3)

Shen YL, Huang CY, Wang SS, Tsao Y, Wang HM, Chi TS (2019) Reinforcement learning based speech enhancement for robust speech recognition. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerrv-Ryan R, et al (2018) Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4779–4783. IEEE

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al (2016) Mastering the game of go with deep neural networks and tree search. nature 529(7587)

Singh SP, Kearns MJ, Litman DJ, Walker MA: Reinforcement learning for spoken dialogue systems. In: Advances in Neural Information Processing Systems (NIPS) (2000)

Singh S, Litman D, Kearns M, Walker M (2002) Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. Journal of Artificial Intelligence Research 16

Sinha A, Akilesh B, Sarkar M, Krishnamurthy B (2019) Attention based natural language grounding by navigating virtual environment. In: IEEE Winter Conference on Applications of Computer Vision (WACV)

Skinner BF (1957) Verbal behavior. new york: appleton-century-crofts. Richard-Amato, P.(1996) 11

Sorokin I, Seleznev A, Pavlov M, Fedorov A, Ignateva A (2015) Deep attention recurrent Q-network. Deep Reinforcement Learning Workshop, NIPS

Steedman MJ (1984) A generative grammar for jazz chord sequences. Music Perception: An Interdisciplinary Journal 2(1)

Strehl AL, Li L, Wiewiora E, Langford J, Littman ML (2006) Pac model-free reinforcement learning. In: International Conference on Machine Learning (ICML)

Su PH, Budzianowski P, Ultes S, Gasic M, Young S (2017) Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. In: Annual SIGdial Meeting on Discourse and Dialogue

Su PH, Gasic M, Mrkšić N, Barahona LMR, Ultes S, Vandyke D, Wen TH, Young S (2016) On-line active reward learning for policy optimisation in spoken dialogue systems. In: Annual Meeting of the Association for Computational Linguistics (ACL)

Su P, Budzianowski P, Ultes S, Gasic M, Young SJ (2017) Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. CoRR abs/1707.00130

Sugiyama H, Meguro T, Minami Y (2012) Preference-learning based inverse reinforcement learning for dialog control. In: Thirteenth Annual Conference of the International Speech Communication Association

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems (NIPS)
Sutton RS, Barto AG et al (1998) Introduction to reinforcement learning, vol. 135. MIT press Cambridge
Takanobu R, Zhu H, Huang M (2019) Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog. In: K. Inui, J. Jiang, V. Ng, X. Wan (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019
Tay Y, Dehghani M, Bahri D, Metzler D (2020) Efficient transformers: A survey. arXiv preprint arXiv:2009.06732
Team T (2021) Code examples: Reinforcement learning. https://keras.io/examples/rl/
Thickstun J, Harchaoui Z, Kakade S (2016) Learning features of music from scratch. arXiv preprint arXiv:1611.09827
Thiemann J, Ito N, Vincent E (2013) The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings. The Journal of the Acoustical Society of America 133(5)
Tjandra A, Sakti S, Nakamura S (2018) Sequence-to-sequence ASR optimization via reinforcement learning. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)
Tjandra A, Sakti S, Nakamura S (2019) End-to-end speech recognition sequence training with reinforcement learning. IEEE Access 7
Ultes Sy, Budzianowski P, Casanueva I, Mrkšić N, Rojas-Barahona L, Su PH, Wen TH, Gašić M, Young S (2017) Domain-independent user satisfaction reward estimation for dialogue policy learning
Ultes S, Barahona LMR., Su PH, Vandyke D, Kim D, Casanueva I, Budzianowski P, Mrkšić N, Wen TH, Gasic M, et al (2017) Pydial: A multi-domain statistical dialogue system toolkit. In: ACL System Demonstrations
Ultes S, Budzianowski P, Casanueva I, Mrksic N, Rojas-Barahona LM, Su P, Wen T, Gasic M, Young SJ (2017) Domain-independent user satisfaction reward estimation for dialogue policy learning. In: F. Lacerda (ed.) Conference of the International Speech Communication Association (INTERSPEECH)
Van Hasselt H, Guez A, Silver D (2016) Deep reinforcement learning with double Q-learning. In: AAAI Conference
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp. 5998–6008
Vezhnevets A, Mnih V, Osindero S, Graves A, Vinyals O, Agapiou J et al (2016) Strategic attentive writer for learning macro-actions. In: Advances in Neural Information Processing Systems (NIPS)
Wang JX, Kurth-Nelson Z, Tirumala D, Soyer H, Leibo JZ, Munos R, Blundell C, Kumaran D, Botvinick M (2016) Learning to reinforcement learn. arXiv preprint arXiv:1611.05763
Wang ZQ, Wang D (2016) A joint training framework for robust automatic speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24(4)
Wang J, Yu LC, Lai KR, Zhang X (2019) Tree-structured regional cnn-lstm model for dimensional sentiment analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28:581–591
Wang R, Ao J, Zhou L, Liu S, Wei Z, Ko T, Li Q, Zhang Y (2021) Multi-view self-attention based transformer for speaker recognition. arXiv preprint arXiv:2110.05036
Wang Z, Bapst V, Heess N, Mnih V, Munos R, Kavukcuoglu K, de Freitas N: Sample efficient actor-critic with experience replay. arXiv preprint arXiv:1611.01224 (2016)
Wang D, Chen J (2018) Supervised speech separation based on deep learning: An overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(10)
Wang Z, Ho S, Cambria E (2020) A review of emotion sensing: Categorization models and algorithms. Multimedia Tools and Applications
Wang Z, Schaul T, Hessel M, Hasselt H, Lanctot M, Freitas N (2016) Dueling network architectures for deep reinforcement learning. In: International Conference on Machine Learning (ICML)
Wang X, Takaki S, Yamagishi J (2018) Autoregressive neural f0 model for statistical parametric speech synthesis. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(8)
Weisz G, Budzianowski P, Su PH, Gašić M (2018) Sample efficient deep reinforcement learning for dialogue systems with large action spaces. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(11)
Weisz G, Budzianowski P, Su P, Gasic M (2018) Sample efficient deep reinforcement learning for dialogue systems with large action spaces. CoRR abs/1802.03753
Whiteson S (2018) TreeQN and ATreeC: Differentiable tree planning for deep reinforcement learning
Williams JD, Raux A, Henderson M (2016) The dialog state tracking challenge series: A review. Dialogue Discourse 7(3)

Williams JD, Zweig G (2016) End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. CoRR arXiv:abs/1606.01269
Williams RJ (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning 8(3-4)
Xin X, Karatzoglou A, Arapakis I, Jose JM (2020) Self-supervised reinforcement learning for recommender systems. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 931–940
Xu L, Zhou Q, Gong K, Liang X, Tang J, Lin L (2019) End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In: AAAI Conference on Artificial Intelligence
Yin H, Pan SJ (2017) Knowledge transfer for deep reinforcement learning with hierarchical experience replay. In: AAAI Conference
Young T, Pandelea V, Poria S, Cambria E (2020) Dialogue systems with audio context. Neurocomputing 388
Yu H, Zhang H, Xu W (2018) Interactive grounded language acquisition and generalization in a 2D world. In: International Conference on Learning Representations
Zamani M, Magg S, Weber C, Wermter S, Fu D (2018) Deep reinforcement learning using compositional representations for performing instructions. Paladyn J. Behav. Robotics 9(1)
Zhang R, Wang Z, Zheng M, Zhao Y, Huang Z (2021) Emotion-sensitive deep dyna-q learning for task-completion dialogue policy learning. Neurocomputing 459:122–130
Zhang Y, Chan W, Jaitly N (2017) Very deep convolutional networks for end-to-end speech recognition. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)
Zhang J, Zhao T, Yu Z (2018) Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog. In: K. Komatani, D.J. Litman, K. Yu, L. Cavedon, M. Nakano, A. Papangelis (eds.) Annual SIGdial Meeting on Discourse and Dialogue, Melbourne, Australia, July 12-14, 2018. ACL
Zhao T, Eskénazi M (2016) Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. CoRR arXiv:abs/1606.02560
Zhao T, Eskenazi M (2016) Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In: Annual Meeting of the Special Interest Group on Discourse and Dialogue
Zhao T, Xie K, Eskénazi M (2019) Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. In: J. Burstein, C. Doran, T. Solorio (eds.) Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)
Zhou H, Huang M, Zhang T, Zhu X, Liu B (2018) Emotional chatting machine: Emotional conversation generation with internal and external memory. In: AAAI Conference on Artificial Intelligence
Zhou Y, Xiong C, Socher R (2018) Improving end-to-end speech recognition with policy learning. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP)
Zhu Y, Mottaghi R, Kolve E, Lim JJ, Gupta A, Fei-Fei L, Farhadi A (2017) Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: IEEE international conference on robotics and automation (ICRA)
Zorrilla AL, Torres MI, Cuayáhuitl H (2021) Audio embeddings help to learn better dialogue policies. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 10.1109/ASRU51503.2021.9688296
Zue VW, Glass JR (2000) Conversational interfaces: advances and challenges. IEEE 88(8)