frontiers | Frontiers in Computer Science

SYSTEMATIC REVIEW
published: 14 July 2022
doi: 10.3389/fcomp.2022.831784

Acoustic-Based Automatic Addressee Detection for Technical Systems: A Review

Ingo Siegert[1]*, *Norman Weißkirchen*[2] *and Andreas Wendemuth*[2]

[1] *Mobile Dialog Systems, Institute for Information Technology and Communications, Otto von Guericke University Magdeburg, Magdeburg, Germany,* [2] *Cognitive Systems Group, Institute for Information Technology and Communications, Otto von Guericke University Magdeburg, Magdeburg, Germany*

Objective: Acoustic addressee detection is a challenge that arises in human group interactions, as well as in interactions with technical systems. The research domain is relatively new, and no structured review is available. Especially due to the recent growth of usage of voice assistants, this topic received increased attention. To allow a natural interaction on the same level as human interactions, many studies focused on the acoustic analyses of speech. The aim of this survey is to give an overview on the different studies and compare them in terms of utilized features, datasets, as well as classification architectures, which has so far been not conducted.

Methods: The survey followed the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) guidelines. We included all studies which were analyzing acoustic and/or acoustic characteristics of speech utterances to automatically detect the addressee. For each study, we describe the used dataset, feature set, classification architecture, performance, and other relevant findings.

Results: 1,581 studies were screened, of which 23 studies met the inclusion criteria. The majority of studies utilized German or English speech corpora. Twenty-six percent of the studies were tested on in-house datasets, where only limited information is available. Nearly 40% of the studies employed hand-crafted feature sets, the other studies mostly rely on *Interspeech ComParE 2013* feature set or *Log-FilterBank Energy* and *Log Energy of Short-Time Fourier Transform* features. 12 out of 23 studies used deep-learning approaches, the other 11 studies used classical machine learning methods. Nine out of 23 studies furthermore employed a classifier fusion.

Conclusion: Speech-based automatic addressee detection is a relatively new research domain. Especially by using vast amounts of material or sophisticated models, device-directed speech is distinguished from non-device-directed speech. Furthermore, a clear distinction between in-house datasets and pre-existing ones can be drawn and a clear trend toward pre-defined larger feature sets (with partly used feature selection methods) is apparent.

Keywords: addressee detection, machine learning, speech-based, multimodal, human computer interaction, systematic review, PRISMA

OPEN ACCESS

Edited by:
Joseph Andrew Allen,
The University of Utah, United States

Reviewed by:
Michelle Cohn,
University of California, Davis, United States
Maria Kovacova,
University of Žilina, Slovakia
Philippe Ravier,
Polytech Orléans, France

***Correspondence:**
Ingo Siegert
ingo.siegert@ovgu.de

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *08 December 2021*
Accepted: *20 June 2022*
Published: *14 July 2022*

Citation:
Siegert I, Weißkirchen N and Wendemuth A (2022) Acoustic-Based Automatic Addressee Detection for Technical Systems: A Review. Front. Comput. Sci. 4:831784. doi: 10.3389/fcomp.2022.831784

1. INTRODUCTION

The general structure of human-machine interaction is based on an easy interface between a human user and a technical system capable of interpreting the commands and information given during an interaction (Sinha et al., 2010). This was traditionally in the form of mechanical switches, like a keyboard, which allowed to convey formerly designated commands to be conferred. The development of more advanced systems leads to the desire to implement more natural user interfaces, allowing for human-like interactions (Biundo and Wendemuth, 2016), ideally without the need to train a user in the knowledge of predefined command lines and instead employing the natural way of speaking as used during a human-human conversation (Siegert et al., 2021).

Additionally, the market for commercial voice assistants has rapidly grown in recent years (Osborne, 2016; Kleinberg, 2018; Kinsella, 2020), mostly because of the ease of use of this technology. Nowadays, this technology has become one of the mainstay products for private household use, with well-known examples like Alexa, Siri, or Cortana from Amazon, Apple, or Microsoft, respectively.

One of the main reasons for this growth is the given naturalness and simplicity of speaking as a form of communication, in contrast to the use of additional external periphery. Simplicity means that the use of a smartphone with an assistant system does not differ substantially from controlling a smart home application. In this sense, a natural interaction is characterized by the understanding (recognition with corresponding feedback) of human expressions and the engagement of people into a dialog, while allowing them to interact seamlessly with each other and the technical environment (Valli, 2007; Baraldi et al., 2009). Furthermore, users don't need to use additional devices or learn any instructions, as the interaction respects typical human communication.

To further improve the naturalness of the interaction, not only improvements in the far-field of speech-to-text interpreter and speech understanding models are needed, but also the ability of a system to discern the addressed interaction partner. In this context, it should be noted that the question of how conversational partners in human-human interaction are addressed has drawn the attention of many researchers (Busso et al., 2007; Gilmartin et al., 2018). Apart from specifying the desirable addressees explicitly by their names or by gaze, implicit acoustic markers are used to emphasize the addressee (Ouchi and Tsuboi, 2016). The latter method is especially useful with regard to special addressees, like hard-of-hearing people (Everts, 2004), elderly people, or children (Garvey and Berninger, 1981; Casillas and Frank, 2017). In the case where the addressee may have some problems of understanding, the normal manner of speech is often self-modified by the speaker.

This modification in the way of speaking is furthermore prevalent in multi-agent environments, which include environments where several users share a technical system, such as smart homes, as well as the opposite situation, where several technical systems are working in the same environment as a human user, such as smart factories (van Turnhout et al., 2005; Vinyals et al., 2012). A situation where both situations may occur simultaneously is of course also possible. Given the ongoing idea of more natural and closely integrated human-machine environments, be it assistant systems, smart home applications, or smart factories, the requirement for such systems to distinguish between system-aimed utterance and independent human-human or even personal exclamation is also growing, to allow for further natural interactions between humans and technical systems (Siegert and Krüger, 2018).

In contrast, the currently preferred interaction initiation regarding voice-based systems is solved in quite an unnatural manner: to detect if a system is addressed, the system is acoustically screening its surroundings waiting for an activation word. The user has to utter the activation word at the beginning of nearly every interaction and wait for the (delayed) activation of the speech-based assistant before starting the request.

Additionally, this method is still error-prone, as the speech-based assistant may not be activated when the wake-word has been said. Sometimes, even worse, it is activated due to a misunderstood acoustically similar phrase (Schönherr et al., 2020; Siegert, 2021), or a phrase utterance (e.g., when the wake-word has been said, but no interaction with the system was intended by the user due to the usage of the wake-word in a different context) (Liptak, 2017; Horcher, 2018).

One main research direction in improving the addressee detection concentrates on the improvement of wake-word detection engines by involving a context-dependent wake-word verification (Wu et al., 2018). But it has already been shown that this method has some weaknesses for phonetically similar utterances (Vaidya et al., 2015; Kumar et al., 2018; Zhang et al., 2019).

This issue is not only disadvantageous due to the adverse usage of the voice assistant, but is also seen as a privacy threat (Chung et al., 2017; Malkin et al., 2019; Dubois et al., 2020), as it leads to unintended recordings during personal interactions. Therefore, voice assistants should be able to perform an addressee detection by themselves, to identify device-directed speech, without the need for an explicit user initiative and more preferably also based on the already gathered utterances. Furthermore, the addressee detection should rely more on acoustic (e.g., prosodic and/or phonetic) information than on a keyword or wake-word detection.

From 2009 onwards, the concept of an automatic speech-based addressee detection system became more prevalent. An increasing number of studies were published, contributing to the field of acoustic-based addressee detection. Especially fueled by the emergence of the first commercially successful smart speakers with Apple Siri in 2011 and Amazon Alexa in 2013. In this regard, the development of natural, secure, and reliable addressee detection methods became imminent. Especially with an increasing employment of integrated human-machine environments, in which systems interpret important information and then communicate their interpretations directly with human observers (Durana et al., 2021; Lăzăroiu et al., 2021), the demand will even raise in the future. As errors in these situations may directly impact the resulting decisions, this could lead to subsequent errors of high impact. Already the interpretation and analysis, as well as the direct control

of human-machine interactions, in economic, healthcare, or transportation relevant environments is dependent on smart processes based on machine-learning and cognitive architectures (Dojchinovski et al., 2019; Gruzauskas et al., 2020; Mahajan et al., 2021; Valaskova et al., 2021; Sri Suvetha et al., 2022). As these methods are employed to reduce the possibilities and impact of economic crisis, an erroneous implementation could be devastating.

1.1. The Early Area of Addressee Detection

But before starting to define the aims and methods of this survey, we will briefly review the beginnings of acoustic addressee detection and present research that has focused on the analysis and characterization of the speech signal to distinguish human-directed speech from device-directed speech.

The underlying difficulty of addressee detection tasks and their problems in a technical system was understood at the beginning of the feasible development of voice-controlled interfaces. The earliest publications (Oppermann et al., 2001; Siepmann et al., 2001), postulated the problems which may arise when a human speaker interacts with a technical system and the system is not able to distinguish between device-directed speech and other utterances. It provided one of the first foundations for the latter research areas. The distinction then was primary toward on- and off-talk. On-talk included in this case all the speech supposed for the system to interpret, while off-talk contains the otherwise spoken parts. This covers not only speech directed at other human participants, but also sudden exclamations, like swearing, reading out loud, or even thinking out loud.

Based on the general behavior during human interaction, another aspect of addressee detection was found in the measurement of gaze giving immediately high and stable addressee information. In Takemae et al. (2004), for example, there was a classification result of roughly 89% in Accuracy (ACC). This could be seen as one of the reasons for the relatively slow and late search for speech-based addressee detectors. This line of research culminated in 2006 into the two directly preliminary developments concerning automated human-human-machine addressee detection. On the one hand, the first gaze-based automated addressee detector was as presented (e.g., in Takemae and Ozawa, 2006). In parallel to this development, there were also in combination with the evolving research in human-human-machine interaction initial analyses of the addressee behavior using human observers (see Lunsford and Oviatt, 2006). The authors provided one of the first estimations on the human classifying ability for different modalities (audio, video, text) and their importance. They concluded that the speaker's and peer's gaze (visual), as well as intonation (acoustic) and, as the authors call it, dialog style (lexical), are most important for the correct judgment. Furthermore, Lunsford and Oviatt (2006) indicated that the annotation was significantly faster based on the audio-only information than on the audio-visual information. Similar conclusions were drawn in Takemae and Ozawa (2006) and Jovanovic et al. (2006) where acoustic and context information was used.

1.2. Definition of Aims and Questions for Survey

To the best of our knowledge, there is no review conducted so far regarding addressee detection in general and acoustic-based addressee detection in particular, which motivates the following survey as the first survey on this emerging topic. Only a few articles comprise a longer review-style state-of-the-art section (cf. Le Maitre and Chetouani, 2013; Siegert et al., 2021), but are neither comprehensive nor discuss the different aspects of an addressee-detection system (data, features, classification architecture, performance) in a comparative manner. By having a longitudinal look at the developments in this research field, it can be shown whether there are particularities that deviate from the general development in related machine learning topics. Therefore, the aim of this review is to identify all relevant studies on acoustic-based addressee detection and discuss similarities and differences in terms of the aforementioned aspects. A further objective of the survey is to identify both promising approaches for future improvements of acoustic-based addressee detection systems as well as gaps in the current research that have to be filled in order to reach natural seamless interaction initiations with technical systems. From these discussions so far, the following aims emerge for this survey. 1) Identify relevant studies that perform an addressee detection based on acoustic information. 2) Compare them regarding datasets, feature sets, and classification architectures. 3) Identify shared methods and important trends.

Although the emerging use of voice assistants encouraged the research on addressee detection, this research is not limited to the use case of voice assistants.

The remainder of this article is structured as follows: In Section 2, we present the structured review guidelines and our criteria. Section 3 presents the procedure to identify the resulting 23 studies and indicated findings regarding used keywords and important conferences. Afterwards, Section 4 discusses the content of the identified studies regarding the research timeline, utilized datasets, features, and classification architectures as well as comparing performance results on studies working on the same datasets. The survey is then concluded in Section 6 giving a final summary and an outlook for future research trends.

2. METHODS

We followed the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) guidelines (Moher et al., 2009) to identify the relevant studies based on eligibility criteria.

2.1. Eligibility Criteria and Literature Search

This survey aims to identify articles that include computational methods for the addressee detection based on or with the primary support of voice characteristics.

The literature search includes all articles that use computational methods based on acoustic characteristics to perform an automatic decision if the utterance was directed toward a technical system or toward a human being. Therefore,

TABLE 1 | Search parameters used in this survey.

Search term	wake-word, wake word, addressee-detection, addressee detection, system-directed, system directed, device-directed, device directed, off-talk, off talk, machine-directed, machine directed, computer-directed, computer directed
Filter term	speech
IEEExplore	("Document Title":**Search term**) OR ("Abstract":**Search term**) OR ("Author Keywords":**Search term**) AND ("Full Text & Metadata":**Filter term**)
Scopus	[TITLE-ABS-KEY (**Search term**) AND ALL (**Filter term**)]
ACM Digital Library	[(Publication Title: **Search term**) OR (Keywords: **Search term**) AND (Keywords: **Filter term**)]
ISCA Archive	**Search term**

it is unnecessary if other additional modalities are used as well (e.g., gaze, lexical information, or dialogue state).

We excluded studies that a) do not distinguish between human-directed and device-directed speech, b) do not base their decision on acoustic speech characteristics, c) do not perform automatic detection (e.g., analyze human decisions), only or (only) do feature comparisons, d) describe acoustic (speech) enhancement methods, e) do rely on lexical (i.e., Automatic Speech Recognition (ASR)-features) or dialog-level features, only, f) only present the task description of an evaluation challenge, g) are position or opinion papers, h) only present or discuss a dataset/corpora for addressee detection, and i) are unpublished or published as non-peer-reviewed material.

As search engines, we used IEEExplore, ACM Digital Library as well as Scopus, as they cover related journals and conference proceedings. Additionally, we searched the ISCA Archive[1], as it includes all conferences, workshops, and symposia hosted by the ISCA from 1987 up to today. Not all of these publications are indexed in the aforementioned indexing services. In contrast to the more elaborated keyword-based search, ISCA Archive only offers a title search.

Regarding the search-term definition, several terms are common to describe the problem of addressee detection, as the field of addressee detection has gained a greater amount of interest together with the increased use of voice assistants. Therefore, we used several search terms with an additional constraint, see **Table 1**. The initial search ran from the 8th of July 2021 until the 16th of July 2021 and was repeated on September, 10th accompanied by the screening of the ISCA Archive, to also include the recent Interspeech 2021 papers.

2.2. Data Extraction

The screening was performed by the two first authors (I.S. and N.W.), by reading the title and abstract as well as the experimental design section. From each study, the used datasets, utilized features, applied classification models, the highest performance or statistical significance, type of validation or test set, and other relevant findings were collected.

[1]International Speech Communication Association (ISCA), https://www.isca-speech.org/archive/#bypaper.

3. RESULTS

A total of 88 studies were assessed for eligibility, after sorting out duplicate records ($n = 145$), records clearly identifiable as unsuitable by the title ($n = 1{,}342$), and records not retrievable ($n = 6$), see **Figure 1**. These records were then assessed for eligibility regarding the exclusion criteria. The following exclusions apply: Description of an acoustic keyword-spotting/wake-word detection without using acoustic information (**reason 1**, $n = 35$), no acoustic information is used (**reason 2**, $n = 10$), no results reported, as only extended abstract or position paper (**reason 3**, $n = 7$), no device-directedness, an only decision between human speakers (**reason 4** $n = 5$), no automatic decision applied (**reason 5**, $n = 7$), and an only description of a database (**reason 6**, $n = 1$). Finally, a total of 23 studies were included in the literature survey.

3.1. Identified Terms in Title and Keywords

As already stated in Section 2, there exist many terms used in literature to describe studies related to automatic addressee detection. The keywords used were addressee detection ($n = 8$), off-talk ($n = 5$), addressability ($n = 1$), focus of attention ($n = 1$), see **Figure 2**. Furthermore, we also analyzed the title regarding addressee detection. Most of the records use "addressee detection" ($n = 12$), "device-directed" ($n = 4$), or off-talk (detection) ($n = 3$) as a phrase in the title. Two studies rely on "system-directed/system-addressed," respectively.

Some records also use a paraphrased description, for example, "talking to a system and oneself," "self-talk discrimination," "talk or not to talk with a computer," or "wake-word-independent verification" ($n = 5$). Two records use a rather general title. This fact has already been seen in the variety of used search terms for this survey, and explains the large difference between identified records and finally included studies.

4. DISCUSSION

4.1. Time Development

To give the reader a visual overview of the chronological order, the 23 identified studies are depicted along a timeline in **Figure 3**.

Groundbreaking work was presented in Batliner et al. (2009), where the authors described a speech-based addressee detection system in its entirety. They specifically distinguished on-talk (system-directed interactions) from off-talk (i.e., interactions toward one-self or other (human) participants).

In Shriberg et al. (2012), a continuation and application of the idea of a system that only listens when a user tries to interact with it was published in 2012. The authors compared different modalities, especially between pure prosodic and lexical features, or a combination of both sets.

A similar approach was taken in Vinyals et al. (2012), where general learning models were researched for streams, including the detection of addressees and speakers. In contrast to a pure speech-based classifier, the system there was optimized for a multimodal approach and included acoustic features only in conjunction with microphone array beams, vision, dialog, situation, turn-taking, and understanding.

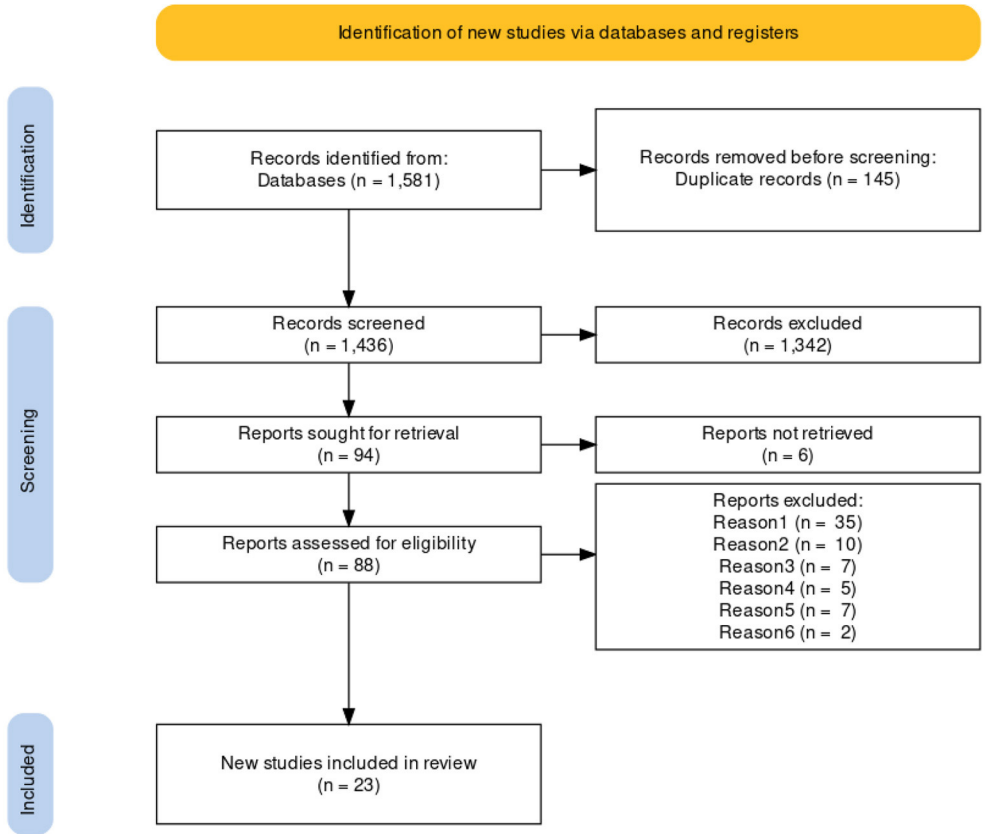

FIGURE 1 | PRISMA flow diagram of study inclusion and exclusion criteria for the systematic review, the detailed reasons are given in the text.

A successor study of Shriberg et al. (2012) was presented in Shriberg et al. (2013), in which further aspects of temporal and spectral features were included in the detection process. Otherwise, the methods and classifier architectures remained the same.

In the same year, a further practical application was tested by Le Maitre and Chetouani (2013), where the primary application was a medical robot in a multiuser environment.

In 2015, two publications, Tsai et al. (2015a) and Tsai et al. (2015b), reported insights on a multimodal and multiuser human interface. Besides typical multimodal features, like acoustic, visual, and lexical data, the authors used dialogue state and beamforming information in addition. Importantly, these studies highlighted that energy-based acoustic features tend to be the most important ones for the distinction of the addressee during an interaction based on their data. Significant care was taken to measure the influence of each used feature set and to devise the optimal combination of them.

Hayakawa et al. (2016a) utilized, for the first time, larger prosodic-acoustic feature sets.

A further development, in the use of prosodic-acoustic features, was made in 2017, where the idea was presented to employ a meta classifier (Akhtiamov et al., 2017a,c). Importantly, the used features were all extracted from speech without an additional visual component this time.

In 2018 and 2019 the research into addressee detection split into several applications and objectives. As an improvement to the previous approach, in 2018 the use of Deep Learning architectures was firstly utilized by Pugachev et al. (2018). Additionally, the relevance of utilizing feature level fusion to incorporate ASR-confidence was closer inspected by Akhtiamov and Palkov (2018). This was necessary, as acoustic features are more complex to train than the semantic or gaze detection methods, but allowing for the independence of domain and language not given by the other two modalities.

In Mallidi et al. (2018) and Huang et al. (2019) the distinction between device-directed and background speech was employed. Also, this research proved that a further subdivision of classifiers for different dialogue types provided much better results than a shared general classifier. In contrast, Norouzian et al. (2019) did not distinguish different sub-groups for the dialogue type and concentrated on the improvement of an attention mechanism. Furthermore, while the presented references so far all used acoustic features independent of wake-words, this is the first approach that specifically posed the idea to remove wake-words altogether from the used dataset. In Akhtiamov et al. (2019) the previous approaches by these authors, such as Akhtiamov et al.

FIGURE 2 | Word clouds as illustration of the used author keywords in the identified studies.

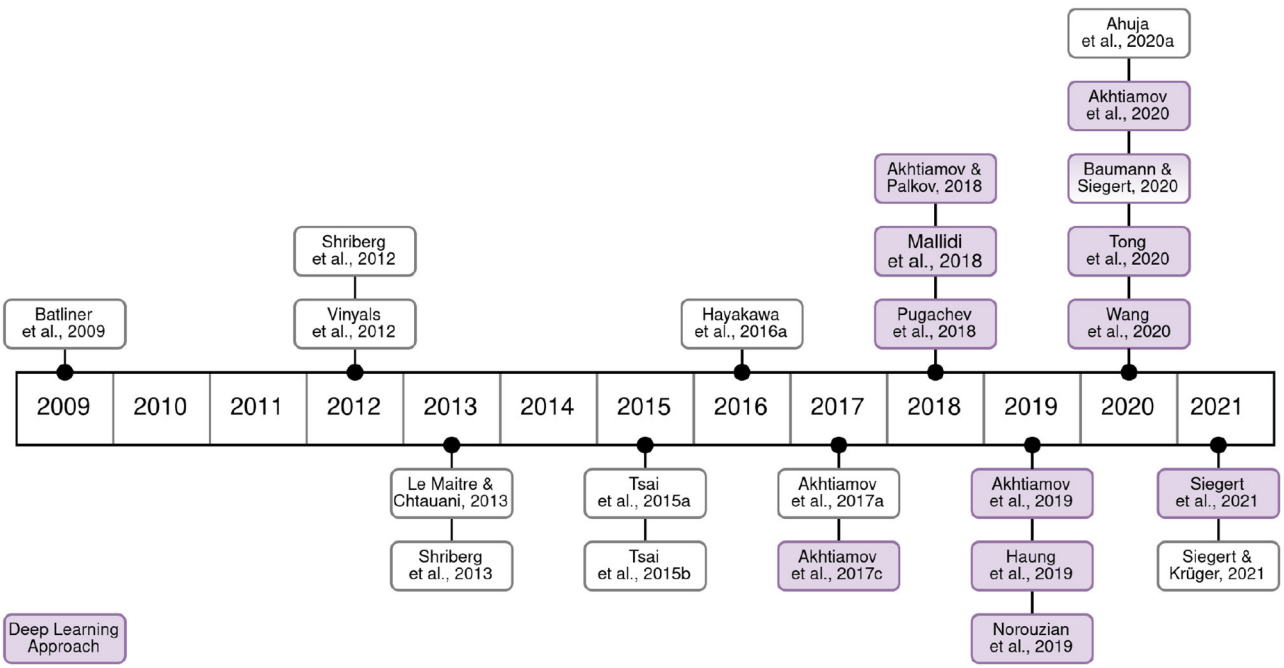

FIGURE 3 | Timeline of the 23 identified studies, studies using deep learning are highlighted.

(2017a) and Akhtiamov and Palkov (2018) were extended with similar methods to before, but were employed on multiple diverse corpora, with the aim to establish a general baseline for further research in this area.

The year 2020 saw a further broadening of important aspects regarding the addressee detection problem. First, with greater importance on true wake-word independence. In Akhtiamov et al. (2020) the classification process was improved by employing

TABLE 2 | Overview of utilized in-house data, sorted by number (#) of utterances.

Year	Study	Data characteristics	Language	# utterances	DD/HD ratio
2013	Le Maitre and Chetouani (2013)	Wizard-of-Oz human-robot interactions	French	516	223:293
2020	Ahuja et al. (2020)	multichannel audio recordings	English	11.5k	N/A
2019	Norouzian et al. (2019)	$\mathcal{D}_1$: Interactions with virtual assistant	N/A	105k	61k:44k
2019	Norouzian et al. (2019)	$\mathcal{D}_2$: As $\mathcal{D}_1$ plus background speech, open microphone recordings, non-speech noise	N/A	134k	7k:127k
2018	Mallidi et al. (2018)	natural interactions with voice-controlled far-field devices	N/A	400k	238k:162k
2019	Huang et al. (2019)	N/A	N/A	324k	N/A
2019	Tong et al. (2021)	natural interactions with voice-controlled far-field devices	N/A	6M	4M:2M
2020	Wang et al. (2020)	true/false wake-words	N/A	"millions"	N/A

Furthermore, datasets from academic research are highlighted. DD, device-directed; HD, human-directed; N/A no information available.

an ensemble classifier, consisting of several classification tasks that are combined in a late fusion approach, which allows combining the strength of the different methods into one singular system. Second, with a heightened sense of privacy by changing a system to ignore information that is not directed to the device either by using features with a limited use to detect what has been said (Baumann and Siegert, 2020) or by extending a wake-word detection system by an acoustic feature classification to improve the security of such a system from false activations (Wang et al., 2020).

Regarding practical applications, Tong et al. (2021) used a combination of the mentioned systems directly during streaming. This work is a continuation of Mallidi et al. (2018) and Huang et al. (2019). In the recent study from 2020, the development went on to use a single input Convolutional Neural Network (CNN) system to exchange the usual feature pre-selection for a system side training process.

A different approach was presented by Ahuja et al. (2020). The authors proposed, similar to gaze detection, to detect the orientation of the speaker. For this, they employed a detection of arrival method, called Direction of Voice (DoV) based on the voice frequency distribution and the crispness, or strength of contrast, of the first wavefront (Ahuja et al., 2020) for details.

In 2021 there was further development into the area of a truly wake-word independent addressee detection system. In Siegert et al. (2021), a system was presented that used a Continuous Learning Framework (CLF) to improve the detection of a specific addressee within the usage of the technical system. By using partial overfitting toward a specific user, better results were archived than with other state-of-the-art methods.

Siegert and Krüger (2021) investigated feature differences between human-directed and device-directed speech, based on self-reported and external reported measures, and afterwards deployed an automatic decision system based on these results. Based on these differences, an automatic system was designed, specifically without wake-word dependency.

In summary, the 23 studies identified all aim to provide acoustic recognition of device-directed utterances, but none of the studies use wake words as features. Wake words as part of the training data set are only used in Huang et al. (2019), the extension of a wake-word detection system by an acoustic-based detection of device-directed utterances is described in Wang et al. (2020).

4.2. Datasets

The datasets used in the reported studies of this survey can be distinguished into two groups. The first group of studies used in-house datasets, that may or may not be publicly available, see **Table 2** for an overview. The second group of studies, used externally available datasets collected within scientific studies, individually published as datasets and available for (academic) research, see **Table 3** for an overview.

4.2.1. In-house Datasets

Among the first group of in-house datasets, the relatively large collections by Nuance and Amazon are especially noteworthy. The size varies between 105 and 134 k utterances for the study by Nuance (Norouzian et al., 2019) and 240 k, 324 k, 6 m utterances, for the studies by Amazon (Mallidi et al., 2018; Huang et al., 2019; Tong et al., 2021). They consist of recordings with virtual assistants or "voice-controlled far-field devices."

They can further include background speech, open microphone recordings, and some non-speech noise. The labels are pre-defined as device-directed or non-device-directed speech. Further details (recording setup, language, speakers, period of recording) are not given.

The other two in-house datasets are recorded within an academic setting, and, as such, of much smaller sample size (Le Maitre and Chetouani, 2013; Ahuja et al., 2020).

Ahuja et al. (2020) recorded their own dataset during the evaluation of their system. It contains 11,520 multichannel English audio recordings of roughly 350 min in length. Recordings of three speakers at polar positions around the device were taken. Le Maitre and Chetouani (2013) recorded an in-house dataset of 543 utterances from eight French-speaking participants during a study on patient-robot interactions. The robot was controlled by a human operator. Each utterance was between 1 and 2.5 s long and labeled into self-talk (293 samples), system-directed speech (223 samples), and unknown (27 samples). In this case, self-talk included also the inter-human-directed utterances.

TABLE 3 | Overview of utilized existing datasets.

Name + Reference	Study	Language	# utterances	DD/HD ratio
HB-CHAAC (Casillas et al., 2017)	Akhtiamov et al., 2019	English	10.8k	6.5k:4.3k
ILMT-s2s corpus (Hayakawa et al., 2016b)	Hayakawa et al., 2016a	English	3.6k	1.2k:2.4k
SmartWeb (Batliner et al., 2009)	Batliner et al., 2009	German	N/A	48.8:51.2%
SmartWeb_acted (Batliner et al., 2009)	Batliner et al., 2009	German	N/A	33.3:66.7%
"Conversational Browser" (Heck et al., 2013)	Shriberg et al., 2012, Shriberg et al., 2013	English	5.5k	2.1k:3.4k
"Trivia-Question Game" (Bohus and Horvitz, 2010)	Vinyals et al., 2012, Tsai et al., 2015b, Tsai et al., 2015a	English	4k	N/A
RBC (Siegert et al., 2019)	Akhtiamov et al., 2020 Baumann and Siegert, 2020 Siegert et al., 2021	German	2.3k	1.5k:0.8k
VACC (Siegert et al., 2018)	Akhtiamov et al., 2019 Akhtiamov et al., 2020 Siegert et al., 2021 Siegert and Krüger, 2021	German	5.6k	3.8k:1.8k
SVC (Batliner et al., 2006)	Batliner et al., 2009 Akhtiamov et al., 2017a Akhtiamov et al., 2017c Akhtiamov and Palkov, 2018 Pugachev et al., 2018 Akhtiamov et al., 2019 Akhtiamov et al., 2020	German	2.5k	1.4k:1.1k

DD, device-directed; HD, human-directed.

4.2.2. Existing Datasets

Within the second group of studies, that rely on existing datasets, we distinguished between datasets used in several studies and used only occasionally and datasets.

A dataset, used already in very early studies of automatic acoustic-based addressee detection and in recent studies as well, is the **Smart Web Video Corpus (SVC)** (Batliner et al., 2006). This dataset was used in seven of the identified studies of this review. This dataset contains 4 h of spontaneous conversations involving 99 German-speaking participants (36 male, 63 female) interacting either with a Wizard-of-Oz (WOZ)-simulated mobile SDS or with a confederate. The participants' age ranges from 15 to 64 years. As context, a visit to the Football World Cup 2006 was used. The user was asking questions of different categories (world cup schedule and statistics, transport, sightseeing, and also open-domain questions), and discussing the obtained information with a confederate, who never talked to the SDS directly. The voice assistant was implemented with a mobile phone as the carrier, similar to a real-world application. The recordings were conducted in public places and, therefore, contain considerable background noise.

Another dataset used in several studies is the **Voice Assistant Conversation Corpus (VACC)** (Siegert et al., 2018). VACC consists of recordings of conversations between a participant, a confederate speaker, and Amazon's ALEXA as a recent commercial voice assistant. This dataset contains 17 h of recordings with 27 German-speaking participants (13 male, 14 female), the age ranges from 20 to 32 years. The participants either have to make appointments with the human confederate speaker or had to answer questions of a quiz, using the voice assistant for support. The recordings took place in a quiet, living room-like surrounding and thus are of high quality.

The **Restaurant Booking Corpus (RBC)** (Siegert et al., 2019) was used in three identified studies. This dataset was explicitly designed to eliminate certain factors influencing the dialog complexity and therefore the addressee behavior of the participant. RBC comprises 90 telephone dialogues of 30 German-speaking students (10 male, 20 female) interacting with either one out of two different technical dialogue systems or with a human interlocutor. The participants' age ranges from 18 to 31 years. The task was to reserve three tables at three different restaurants for four persons under certain constraints (late dinner time for 1 day, sitting outside, reachable with the public transport system, availability of vegetarian food, receiving directions to the restaurant).

A dataset captured at Microsoft-Research recorded several groups of two participants and a dialogue system, which resulting dataset is denoted as the **"Trivia-Question Game"** (Bohus and Horvitz, 2010). The scenario involves groups of two to three people playing a trivia question game with a computer agent. This dataset consists of a total of 148 interactions, separated into 89 two-party and 59 three-party examples, all in English. In total, there were 4,605 spoken utterances detected by the system, which were manually annotated with source and addressee information.

Another dataset, also collected by researchers at Microsoft is the "**Conversational Browser"** (Heck et al., 2013). The Browser was installed visibly as a large TV screen about 5 feet (1.52 meters) away from the seated participants, which were roughly 3 feet (0.91 meters) away from each other. Using a set of predefined commands, participants could start a new interaction, pause, stop listening, or "wake up" the system. The full corpus comprised 6.3 h of English speech recordings and was taken from 17 different groupings of 13 unique speakers. The sessions ranged from 5 to 40 min. The speech was captured by a Kinect microphone; end pointing and recognition used an off-the-shelf recognizer.

Batliner et al. (2009) presented two additional databases, **SmartWeb_acted** and **SmartWeb_spont**. The prompted speech utterances in **SmartWeb_acted** are intended to research which classification rates can be achieved and to show the differences to real data. The content of the acted data is based on the same scenario as the SVC corpus. This dataset comprises 17 German speakers, the gender distribution is unclear. It contains pre-formulated sentences toward a fictive system and fictive dialog partners and speakers were given detailed instructions on how

to pronounce On-Talk and Off-Talk. This set has a total amount of 1.7 h of recordings. In **SmartWeb_spont**, 0.8 h of data from 28 German speakers have been utilized. In this public scenario, the user can get information on specific points of interest, e.g., hotels, restaurants, or cinemas. The system is employed as a WOZ experiment and non-prompted, spontaneous speech utterances are gathered.

A dataset only used in one of the identified studies (cf. Hayakawa et al., 2016a) is the **ILMT-s2s corpus** (Hayakawa et al., 2016b). It represents a multimodal interlingual Map Task Corpus collected at Trinity College, Dublin. In total 15 dialogues (9.5 h) were collected during the map task setting where the instruction giver and the instruction follower speak different languages (English and Portuguese) and are mediated by a speech-to-speech translation system used. The corpus, containing English speech, comprises different data streams: audio and video recordings of the subjects, as well as heart rate skin conductance, blood volume pressure, and electroencephalography signals of one subject in each dialogue. Afterwards, manual annotation was conducted according to different schemes, including On-Talk and Off-Talk (self and other). This differentiation into On-Talk (talking to the translation system and thus, in consequence, to the other user, 2,439 utterances) and Not-talking to that system (1,189 samples) is then used in the reported addressee detection studies by Hayakawa et al. (2016a).

Another dataset, only used in one of the identified studies (Akhtiamov et al., 2019) and not directly intended for human vs. device addressee detection, is the **HomeBank Child-Adult Addressee Corpus (HB-CHAAC)** (Casillas et al., 2017). This dataset was part of the INTERSPEECH 2017 Computational Paralinguistic Challenge: Addressee, Cold & Snoring (Schuller et al., 2017) and comprises 10,861 manually annotated speech utterances of parents, interacting with each other and their children (0–20 months), in their home. To capture the speech signal, a LENA audio recorder, worn by the children in specialized clothing, was used[2]. This dataset was used in Akhtiamov et al. (2019) together with two datasets comprising human-directed and device-directed utterances (SVC and VACC), with the authors' assumption that device-directed speech is prosodically similar to children-directed speech present in HB-CHAAC. However, studies suggest that the pitch range between speech directed at children and speech directed toward computers differs (smaller for computers, larger for children) (Mayo et al., 2012). This observation is also supported by the experiments of Akhtiamov et al. (2019), as the incorporation of HB-CHAAC does not improve the recognition performance and the authors state that in cross-corpus experiments the neural networks tend to ignore the information from the HB-CHAAC data.

4.2.3. Studies Including Several Datasets

Some identified studies in this review also used multiple datasets, either in terms of increasing the amount of training material or testing the generalizability of trained models.

Norouzian et al. (2019) are using two datasets in terms of cross-corpus tests, specifically two in-house sets, with and without an open-mic setup in a noisy environment. The models trained on $\mathcal{D}_1$ are afterwards tested on $\mathcal{D}_2$. With their final model architecture, comprising a CNN layer and a Long Short-Term Memory (LSTM) layer with attention, the decrease in Equal Error Rate (EER) is only about 0.63% absolute, and the generalizability is reportedly high.

Some studies, conducted in 2019 and 2020 by Akhtiamov et al. and Siegert et al., combined several datasets to increase the amount of training material. They utilized SVC, VACC, and RBC as well as mixup data augmentation (Zhang et al., 2018). For using RBC as a test set, one of the best performances of 60.90% Unweighted Average Recall (UAR) was achieved, when using VACC data together with RBC data with an *end-to-end* (*e2e*) speech processing model (Akhtiamov et al., 2019). But using a more complex meta-model, that makes use of different models to combine different layers of information, gives a slightly better performance of 62.80 UAR (Akhtiamov et al., 2019, 2020; Siegert et al., 2021).

Additionally, some studies also investigated cross-corpus experiments using SVC and VACC (Akhtiamov et al., 2019) and in combination with RBC (Akhtiamov et al., 2020). The authors used all of these three datasets for testing and training. Regarding their experiments, they revealed a clear similarity between VACC and SVC.

4.2.4. Summary–The Struggle of Availability and Amount of Material

Foremost, it is noteworthy that all studies, except Batliner et al. (2009), fully relied on realistic data from the beginning and did not just use acted data. Only a minority of academic studies used an in-house dataset, which allows other researchers to reproduce the results and test further approaches for improvement. Furthermore, these datasets are quite small and designed for a specific purpose, impeding the transfer to other domains.

Due to the rise of commercial voice assistants, efforts by industrial research also increases in this area, often using their own in-house data collections. In contrast to the academic in-house data, these sets comprise orders of magnitude more language material. Unfortunately, industrial approaches are not tested against public available corpora and there is a comparatively low amount of open and public available data. As commercial datasets also often contain personal identification information, a free exchange between these two sources is seldom possible.

Studies using corpora regarding automatic speech-based addressee detection rely on English and German datasets, all recorded to either study human-human-computer interaction in general or are specifically designed to analyze the addressee behavior in group settings, with one technical counterpart. In contrast to the previously mentioned in-house datasets, they comprise much less material.

Regarding the identified studies and although several datasets are employed, it can be stated that the variations

[2]We are aware that the task of distinguishing child vs. adult addressee is not the same as human vs. device addressee, but for the sake of completeness we intended to mention this dataset.

TABLE 4 | Overview of utilized features/feature sets for each identified study.

Year	Study	Utilized feature/Feature set	Size*
2013	Le Maitre and Chetouani, 2013	F_0 (8), energy (8), rhythm (3)	19
2012	Shriberg et al., 2012	energy peak related (10), speech activity (4), energy contour (7)	21
2020	Ahuja et al., 2020	high-low band ratio (9), crispness of first wavefront (23)	32
2013	Shriberg et al., 2013	Energy contour (7), voice quality and spectral tilt (3), delta energy at voicing onsets/offsets (14)	24
2020	Baumann and Siegert, 2020	*MFCC* (13), *FFV* (13), phonetic segment network (10)	36
2015	Tsai et al., 2015b	energy (21), energy change (24), temporal shape of speech energy contour (2)	47
2015	Tsai et al., 2015a	energy (21), energy change (24), temporal shape of speech energy contour (2)	47
2020	Wang et al., 2020	*LFBE*	64
2018	Mallidi et al., 2018	*LFBE*, 1-best ASR-hypothesis, ASR decoder features (3)	68
2009	Batliner et al., 2009	duration, energy, *Fundamental Frequency* (F_0), pauses, jitter, and shimmer	100
2012	Vinyals et al., 2012	amplitude, average log energy per utterance/at different regions of the utterance	128
2020	Tong et al., 2021	*log-STFT256*	256
2019	Huang et al., 2019	*LFBE* (64), *log-STFT256*, ASR decoder (4)	260
2021	Siegert and Krüger, 2021	*emobase* (988)	988
2017	Akhtiamov et al., 2017c	*ComParE* via recursive feature elimination	1,000
2017	Akhtiamov et al., 2017a	*ComParE* via recursive feature elimination	1,000
2018	Akhtiamov and Palkov, 2018	*ComParE* via recursive feature elimination	1,000
2016	Hayakawa et al., 2016a	*ComParE* via recursive feature elimination	6,256
2018	Pugachev et al., 2018	*ComParE*	6,373
2019	Akhtiamov et al., 2019	*ComParE*, e2e	6,373
2019	Norouzian et al., 2019	Log mel-filterbank (45), *ComParE*	6,373
2021	Siegert et al., 2021	*emobase* (988), *ComParE* (6,373), ASR-confidence (1), e2e	6,373
2020	Akhtiamov et al., 2020	*ComParE* (6,373), ASR-confidence (1), e2e	6,374

*Sorted according to the largest number of features used. Studies using widely known feature sets are highlighted. *Size stands for the number of features used.*

regarding speaker characteristics (age, gender, region, race) and language characteristics (language, dialect, slang) is too limited. Thus, additional research is needed to investigate the performance of addressee detection regarding broader variations. Further research is needed, especially regarding under-resourced languages (Besacier et al., 2014) and, as recent discussions have shown, regarding all types of social and racial disparities (Koenecke et al., 2020; Martin and Tang, 2020).

4.3. Features

Regarding the identified studies in this survey, the utilized features can be distinguished into two groups: I) smaller hand-crafted feature sets and II) widely known (larger) feature sets. Especially the early studies, done before 2016 (Shriberg et al., 2012; Vinyals et al., 2012; Le Maitre and Chetouani, 2013; Tsai et al., 2015b), used manually selected features, as the discriminative power of specific features was unknown and the use of smaller feature sets helped in the training phase, see **Table 4** for a broad overview. Regarding an in-depth understanding of the mentioned features, the reader is referred to the Springer Handbook of Speech Processing (Benesty et al., 2008) and the openSMILE book (Eyben et al., 2013).

4.3.1. Studies Using Hand-Crafted Feature Sets

In Batliner et al. (2009) the most prevalent (speech) features were duration, energy, pitch (i.e., F_0), pauses, jitter, and shimmer[3]. These raw features are described in several derivatives including, for example, temporal developments, and absolute or mean values. In total, this study used 100 acoustic features. The primarily used features in Shriberg et al. (2012) were energy and speaking rate, extracted at the level of a Kinect segment[4]. Also included were peak-dependent features, auch as peak count, rate, and several distance measurements between each peak. Additionally, the set contains speaking rate and duration information, as well as waveform duration and length information based on speech activity information[5]. Another used prosodic-acoustic feature set was energy contours, these consisted of *Mel-Frequency Cepstral Coefficients* (*MFCCs*) which various sliding windows as well as the first five *Discrete Cosine Transform* (*DCT*) basis functions. Importantly, all acoustic features were chosen to be word, context, and speaker-independent.

[3]Jitter and shimmer represent variations on the fundamental frequency. Jitter refers to irregularities in the fundamental frequency or period of a speech signal, mainly because of lack of control of vocal fold vibration. Shimmer refers to the superposition of the fundamental frequency of a speech signal with a noise, so that irregularities occur in the amplitude due to reduction of glottic resistance and mass lesions in the vocal fold (Haji et al., 1986).

[4]A segment according to the authors is a complete utterance as detected by the Kinect with an "off-the-shelf" recognizer.

[5] Shriberg et al. (2012) computed these features from the time-alignment of the word recognition output within the region that triggered speech activity detection, without making reference to the identity of the recognized words.

In addition to the features from Shriberg et al. (2012), such as energy contour, further information was gained in Shriberg et al. (2013) from voice quality and spectral tilt characteristics. This was done because of a perceived higher "vocal effort" during device-directed speech. This same effort was also detected by the delta values of energy during the onset and offset of the voicing.

Based on similar earlier approaches, Le Maitre and Chetouani (2013) also used F_0 as well as the typical energy-based features. Both sets contained the temporal derivations in addition to the values based on segments (i.e., maximum, minimum, mean, standard deviation, interquartile range, mean absolute deviation, and quantiles), which led to 16-dimensional vectors. Additionally, rhythmic features were extracted by using perceptual filters searching for prominent events. With these the mean frequency, entropy and barycenter were determined.

The features used in Tsai et al. (2015a,b) were taken from a variety of sources, these were the acoustic, ASR, system, visual and beamforming sets. Specifically, the acoustic features were combined from energy features (i.e., energy levels), during different time segments of an utterance. Derived from this were the energy change features and the general energy shape contour. The similar-sized visual feature set consisted of movement, face orientation, and physical distance between the participants. The additional features were the general dialogue state, the beamforming direction, and the lexical features, such as n-grams and speech recognition. In total 117 features are available, with 47 of them from the acoustic family.

The experiments in Baumann and Siegert (2020) reduced the used feature set to *MFCC*s as well as *Fundamental Frequency Variation*s (*FFV*s). This was partially done because the extractable information from these features does not allow for an easy reconstruction of the spoken message, improving the security and privacy of the approach.

4.3.2. Studies Using Widely Known Feature Sets

Features used in Akhtiamov et al. (2017a,c) and Akhtiamov and Palkov (2018) could be distinguished into parsed speech, acoustic features, and textual content, which could be further distinguished into the syntactical and lexical analysis. The acoustic speech analysis was based on the former generated knowledge of general louder and more rhythmical speech patterns in the case of device-directed speech, which aims to make the speech by the human speaker easier to understand for a technical system. To cover a broad set of possible characteristics, Akhtiamov et al. (2017a,c) and Akhtiamov and Palkov (2018) have chosen the *Interspeech ComParE 2013* (*ComParE*) feature set (Schuller et al., 2013). It provides 6373 different features containing for example basis energy, spectral, voicing related descriptors, and the often employed *MFCC* features. Additionally, descriptors such as logarithmic harmonic-to-noise ratios, spectral harmonicity, and psychoacoustics-related spectral sharpness were included. To reduce the redundancy and improve the classifier result, the authors employed a recursive feature elimination model, for gaining more significant information. With the help of a Support Vector Machine (SVM), the relative weight of each feature was defined and sorted by its importance, details on the exact procedure on how to perform this feature selection are given in Akhtiamov et al. (2017c). One thousand features were chosen based on this. The text analysis comprises, for example, recognition confidence, number of recognized words, and utterance length as well as part-of-speech n-gram.

Pugachev et al. (2018) compared both strategies and used both the total number of 6,373 features of *ComParE* and a reduced set of 1,000 features utilizing the above-explained approach.

The same total feature set was also used in Hayakawa et al. (2016a). K-Means was used to define the most important features, which lead to a small reduction from 6,373 to 6,356 features. In Akhtiamov et al. (2019), the baseline generating publication, the *ComParE* feature set was used as well, in conjunction with a recursive feature elimination. This showed a complicated aspect of the addressee detection, as each examined dataset generated another set of representative features. Specifically, the study identified 450 relevant features for VACC, 2020 for SVC, and 400 for HB-CHAAC. The essential greater size for SVC is attributed to the Wizard-of-Oz design of the dataset. Furthermore, only 28 identical features were identified among all three datasets[6]. When only comparing VACC and SVC, which are designed with the same target classes, this rises to 172 similar features. In Akhtiamov et al. (2020) where a meta-classifier approach is used, again the *ComParE* feature set is used, together with the ASR information and the spectrogram representation for the e2e approach.

In addition to the previously mentioned feature sets, Siegert et al. (2021) also used *Opensmile's emobase* (*emobase*) set comprising 988 features. This set contains similar features to the automatically reduced *ComParE* feature set, *emobase* is also used in Siegert and Krüger (2021).

In Norouzian et al. (2019), a spectrogram image of the log-mel filterbank features is employed, to train a CNN architecture. This reduces the necessary preparation steps, as the data does not need to be prepared beforehand. The parallel introduced utterance-level model uses the typical *ComParE* feature set, with a total number of 6,373 features.

Wang et al. (2020) used CNNs for their experiment, too. To therefore enable a visual representation, the authors have utilized *Log-FilterBank Energy* (*LFBE*) features, extracted before and after the use of a wake-word.

In Mallidi et al. (2018), *LFBE* as acoustic embeddings are utilized together with the 1-best hypothesis from the ASR decoder and additional features from the ASR decoder (e.g., trellis, Viterbi costs). *LFBE*, as well as ASR decoder features, are also used in Huang et al. (2019). Additionally, the authors also chose features representing the *Log Energy of Short-Time Fourier Transform* (*log-STFT256*).

In Tong et al. (2021) the *log-STFT256* features are used as a visual representation for a CNN classifier, as well. As this research was a continuation of the former primarily multimodal-based approach (consisting of ASR, embeddings, and decoder features)

[6]These features are different functionals of the following low level descriptors and their deltas: F0final_sma, audSpec Rfilt_sma, mfcc_sma, pcm_fftMag_spectralRollOff25.0_sma, pcm_fftMag_spec_tralRollOff50.0_sma, voicingFinalUnclipped_sma.

this showed a preference for the system to train their own feature extraction instead.

Furthermore, three studies employed an e2e classification path, directly working on the acoustic representation (Akhtiamov et al., 2019, 2020; Siegert et al., 2021).

4.3.3. Summary–General Trends on Feature Sets

Summarizing the section about the different employed features, it can be stated that the studies follow the overall movement in acoustic analyzes from hand-crafted, small feature sets, like loudness, F_0, and duration toward more standardized (widely-known) feature sets with separate trends for both broader and specialized feature sets. The identified studies from 2009 toward 2013 utilized different hand-crafted features with energy information as an integral part (Batliner et al., 2009; Shriberg et al., 2012, 2013; Vinyals et al., 2012; Le Maitre and Chetouani, 2013), with a short resurgence in Tsai et al. (2015a,b). Beginning with Hayakawa et al. (2016a), the usage of full feature sets became usual, specifically the *ComParE* feature set from the Interspeech 2013 challenge (Schuller et al., 2013). This set was employed in Hayakawa et al. (2016a), Akhtiamov et al. (2017a, 2019, 2020), Akhtiamov and Palkov (2018), and Pugachev et al. (2018). To still reduce the necessary computational power, this set was often optimized, either by feature elimination/selection or by machine learning pre-training with an attention mechanism. A set of features comparable to *ComParE* is *emobase*, as used partially in Siegert et al. (2021).

In parallel, with the development of deep learning networks and specifically CNNs, in Huang et al. (2019), Norouzian et al. (2019), Ahuja et al. (2020), Wang et al. (2020), and Tong et al. (2021) *LFBE* and *log-STFT256* features were utilized. This was partly also because of the feature extraction ability from visualized data, inert in CNNs. Compared to this, Baumann and Siegert (2020) specialized features that emphasizes privacy concerns were employed. The chosen *FFV* is of limited use for recognizing what is spoken, as it features only prosodic information.

A discussion regarding the influence of the number of features and the type of features on the addressee detection performance is given in Section 4.4.5.

4.4. Classification Architectures and Reported Performances

4.4.1. Short Overview of Utilized Performance Measures

As the identified studies originate from different research fields, different performance measures are used to report the classification result. For a more detailed explanation, the reader is referred to Olson and Delen (2008), Powers (2011), and Siegert (2015).

In terms of optimizing the recognition system to have equal performance in falsely accepting a human-directed utterance as device directed or falsely rejecting a device-directed utterance, the Equal Error Rate (EER) is used. The lower the EER, the better the system.

In terms of indicating the correct detection of a class, the most commonly used evaluation measure is the Accuracy (ACC). It measures the percentage of correct predictions. Higher accuracies denote better performance. But this measure provides no statements about failed classifications, which must be taken into account to compare the results of different classifications.

Thus, to estimate the effectiveness or completeness of a classification, the sensitivity (e.g., recall or true positive rate) is used. It measures the proportion of all samples correctly classified and the total number of samples in the class. In the case of an addressee-detection problem, it is usually of interest to measure the recall of both classes, i.e., device-directed and human-directed. Therefore, the Unweighted Average Recall (UAR) is used, by averaging the summarized class-wise recall over the number of classes. A higher number indicates a better classifier.

Sometimes, it is also important to indicate the ratio of truly correct samples within the total number of samples classified as correct, in other terms, to denote that the number of false classifications is small. This can be expressed by reporting the precision additionally to the recall. As it is not possible to optimize all measures simultaneously, combined measures are used to have a single value judging the quality of a classification. One commonly used measure is the F-measure. It combines precision and recall using the harmonic mean.

All of these measures are used for good reasons, but without the individual recognition results for the single classes or the recognition scores they cannot be transferred into each other.

4.4.2. Classical Classification Architectures

The mid-2010s studies used classical approaches, especially as the amount of training material was limited. The employment of different classical classification architectures within the identified studies is depicted in **Figure 4**.

One of the first used classical classifiers for the automated addressee detection were Linear Discriminant Classifiers used in Batliner et al. (2009). The authors used a typical leave-one-speaker out training method. With only acoustic features, the system achieved an ACC of 74.2% for SVC and 66.8% for Smartweb. With the addition of part-of-speech features, the authors achieved an ACC of 74.1 and 68.1%, respectively. Only considering the acted part of SmartWeb, results of 92.6% (ACC) were possible when using prosodic speech features with speaker normalization.

Even with vast multimodal inputs, as in Vinyals et al. (2012), only an EER of 10.8% and 13.9% were achieved, either with random forest or maximum entropy architectures. As the different features were not observed individually a clear influence is not possible, but it mirrors the other available results from around this year (e.g., Shriberg et al., 2012, 2013).

The study by Le Maitre and Chetouani (2013) used a variety of different classifiers: decision trees, k-NN, and SVMs with a radial basis function. Given the used dataset, a 10-fold cross-validation was used instead of a leave-one-speaker-out scheme. The resulting accuracy was comparably low with 55.46% for decision trees, 58.20% for k-NN, and 71.62% for SVMs. It depended on its full feature set, as it fell below 60 % when only one feature aspect was employed, such as only energy, rhythm, etc.

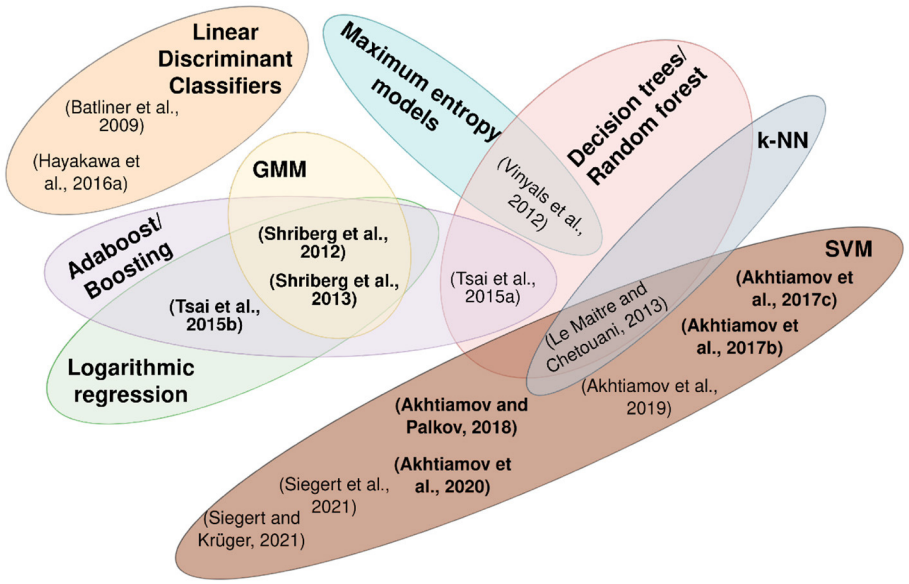

FIGURE 4 | Venn Diagram to illustrate the use of different classical classification architectures in the identified studies. Studies in bold employ a meta-classifier.

In the work of Tsai et al. (2015b) Adaboost with tree stumps is used, which is extended in Tsai et al. (2015a), where specifically logarithmic regressions, decision trees, random forests, and Adaboost with tree stumps are utilized. Each of them proved to be capable of classifying the general addressee detection with decision trees only reaching around 17% EER while random forest, logarithmic regression, and Adaboost achieved all-around 14% EER.

Discriminant analysis was used in Hayakawa et al. (2016a). This method assumes that different cases of on-talk and off-talk were part of different Gaussian distributions. The averaged F1-score achieved with this method is 90.85%. The authors employed further experiments to distinguish on-talk from un-directed off-talk (e.g., self-talk) and off-talk directed to another person. The average F1-score, in this case, was only 69.41%, where especially the un-directed off-talk showed a dramatically low recognition performance (36.64%). The classifier used in Siegert et al. (2021) is a continuously trained SVM architecture. By applying a pre-training step, a relatively simple classifier that is continuously improved with each further interaction between the system and the user could be optimized. By using the stepwise adaptation, the final results of the system were higher than the results from the more complex meta classifier as used in Akhtiamov et al. (2020). For the used RBC datasets human listeners achieved a UAR of 53.57%, a simple linear SVMs achieved 52.02%, the complex meta-classifier archived 52.70% and the continuous learning framework outperformed the previous approaches with 85.77% UAR.

4.4.3. Deep Learning Architectures

Beginning with 2018, the increased emergence of larger datasets, higher-performance algorithms, and the general popularity of the use of deep learning approaches for addressee detection also raised. The identified studies making use of deep learning approaches are depicted in **Figure 5**.

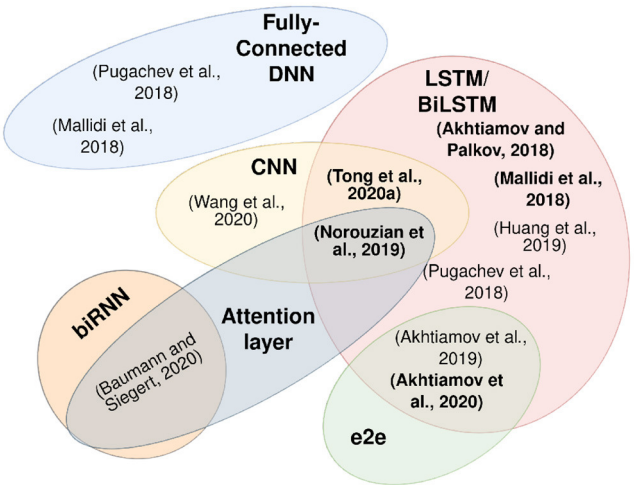

FIGURE 5 | Venn Diagram to illustrate the use of different deep-learning classification architectures in the identified studies. Studies in bold employ meta-classifier.

The first deep learning architecture was then employed in Pugachev et al. (2018) and consisted of a two hidden layered, fully-connected feed-forward network. This was used in conjunction with the reduced dataset of the 2017 experiments (Akhtiamov et al., 2017b,c), using acoustic features alone. By adopting several sizes of hidden layers and optimization methods, the classification performance could be tuned for better performance. The final system showed a UAR of 78% with an

F1-score of 78% and an ACC of 69%. Furthermore, the authors also employed a Bidirectional-LSTM (biLSTM) as an additional method. These provide the ability of a recurrent neural network with a deep layer architecture, but also require a greater amount of training data. This resulted in a Deep Neural Network (DNN) system achieving the best performance with an ACC of 78%, in comparison to the original SVM with only 70%. The data dependency of the biLSTM results in a decreased performance of only 65% ACC. The result was then further examined and improved by Akhtiamov and Palkov (2018), which worked on a multimodal approach with a meta classifier architecture. The system itself allowed the partial usage of only certain feature sets and achieved a UAR of 80% on the acoustic set only.

In Mallidi et al. (2018) two LSTMs are trained to either capture the acoustic embeddings or the ASR 1-best hypotheses. With this approach, the authors achieved an EER of 9.3, 10.9, and 20.1%, for acoustic embeddings, and 1-best embeddings, respectively. In Huang et al. (2019) a slightly changed approach is used. An LSTM was trained directly on the frame-wise representation of the recordings, with the used acoustic embeddings extracted specifically from the pre-softmax output. With these features, the effect of the pure decoder was enhanced compared to older methods by reducing the relative EER to about 29.54% with the *LFBE* features and about 32.39% with *log-STFT256*[7]. This also improved the relative ACC by about 6.31 and 6.92%, respectively. In absolute numbers, the assumed EER is around 3.66% for *LFBE* and 3.52% for *log-STFT256*[8].

In Akhtiamov et al. (2019) three different classifiers are tested against each other. The first, is a typical SVM architecture with a linear kernel, the second a two stacked LSTM with a global max pooling, dropout, and softmax layer, and the third is an end-to-end architecture. Because of architectural differences, the second classifier receives frame-wise input for the recurrent capabilities of the LSTM, while the end-to-end classifier receives its information from the output of a CNN, which in turn feeds from the original signal. The SVM achieved a UAR of 78.8%, the LSTM of nearly 90.0%, and the e2e method of 85.3%. Each of these results was gained on VACC, the compared datasets of SVC and HB achieved worse results. With a greater emphasis on the recurrent capabilities, which can continuously feed information into the system, Baumann and Siegert (2020) used a combination of a bi-directional (recurrent-) neural network (biRNN) with an added attention functionality and a hidden layer with softmax classifier. Additionally, the study provided information on the general capabilities of a human listener to classify the same data, which was between 53.5 and 60.6% depending on the language of the listeners and the samples. The automatic classifier achieved F1-score results of 65.5% for all features and 63.8 or 63.3% for *FFV* or *MFCC* only, respectively.

The CNN of Wang et al. (2020) consists of two independent networks of five layers, each either using the information from before or after the wake-word usage. The results of both are then fed into a three-layer densely-connected feed-forward network, which then feeds the prediction set. This system was not tested without a wake-word, but improved the capabilities of such a system without the need for much greater training data.

4.4.4. Classifier Fusion

The classifier fusion has been used in several studies, where the output of different classifiers, using the same input vectors or different information streams, are combined using an additional classification layer or classifier to improve the overall classification performance (see Lalanne et al., 2009; Glodek et al., 2015) for an overview of fusion techniques.

One of the earliest works using a classifier fusion approach are the works of Shriberg et al.. In Shriberg et al. (2012) the prosodic features were trained with Boosting for the segment-level features, and Gaussian Mixture Models (GMMs) were used for the energy features. For the combination of both classifier outputs, a Linear Logistic Regression (LLR) was then utilized. Boosting on segment-level features achieved an EER of 16.03% and GMM with energy features achieved an EER of 13.93%, the combination using LLR achieved an EER of 12.63%. In Shriberg et al. (2013) the same classifiers were employed and with additional features. An EER of 12.50% could be achieved at best for the LLR fusion.

In Akhtiamov et al. (2017a,c), and Akhtiamov and Palkov (2018) the idea of a multimodal approach was optimized not only by adapting the used classifier but also by employing an adapted meta-classifier on top of the acoustic and textual classification. This led to several layers of classifiers working in parallel. The acoustic and textual features were combined on a feature level before using a linear SVM, while different SVMs were applied to the textual aspects alone (specifically the stemming and POS tagger parts). The results were then combined by a linear SVM meta-classifier. By removing aspects of this architecture, different parts could be observed concerning their performance. The authors presented different parameter tuning and feature selection approaches in the three papers, where the best performance with only the acoustic information of 82.2% (UAR) and 82.8% from acoustic and ASR-information, could be achieved (Akhtiamov et al., 2017a,c). This single-modality performance could be optimized with other aspects of the meta-classifier to 92.9% (UAR) using acoustic, syntactical, and lexical information. In Akhtiamov and Palkov (2018) some slight optimization in the date representation was examined, but neither the acoustic-only classifier, achieving 82.0% UAR nor the meta-classifier combining visual, acoustical, and textual information achieving 92.6% UAR could outperform the previous results.

A similar meta-classifier is used in Akhtiamov et al. (2020) and achieved a UAR of 84.9% on the VACC dataset. The classification system is a combination of a *ComParE* functionals' classifier, using a linear SVM, a *ComParE* LLD classifier, using a biLSTM with deep learning feedforward network at the end, an ASR classifier with Meta-feature extraction, and an e2e classifier using the visual representation with a CNN feeding into a biLSTM comparable to the LLD classifier. The results, with their confidence score, are then concatenated in a linear SVM for the final prediction.

[7]The authors just reported relative improvements in comparison to older papers of the same group.

[8]As calculated in this survey by comparing the mentioned references in the paper.

TABLE 5 | Comparison of DD/HD recognition performance of selected identified studies, compiled regarding the utilize dataset.

Reference	Measurement	Value[%]	Features, Classifier
"Conversational Browser"			
Shriberg et al. (2012)	EER	12.63	GMM (energy features), Boosting (energy + speaking rate features), combined with LLR
Shriberg et al. (2013)	EER	12.50	GMM and Boosting (energy contour, spectral tilt and delta energy at voicing onsets/offsets), combined with LLR onsets/offsets (14)
"Trivia-Question Game"			
Tsai et al. (2015a)	EER	16.39	Adaboost on all 177 features, including 47 acoustic one
Tsai et al. (2015b)	EER	13.90	Adaboost on all 177 features, including 47 acoustic one
Vinyals et al. (2012)	EER	10.80	Expert random forest model, manually crafted features
SVC			
Batliner et al. (2009)	ACC	74.20	Linear Discriminant Classifier on 100 prosodic features
Pugachev et al. (2018)	UAR	78.00	DNN (1,000 *ComParE* features)
Akhtiamov and Palkov (2018)	UAR	80.00	SVM (1,000 *ComParE* features with recursive feature elimination and ASR-confidence)
Akhtiamov et al. (2017a,c)	UAR	82.20	SVM (1,000 *ComParE* features with recursive feature elimination)
Akhtiamov et al. (2017a,c)	UAR	82.80	SVM (1,000 *ComParE* features with recursive feature elimination and ASR-based features)
Amazon in-house dataset, using Mallidi et al. (2018) as baseline			
Mallidi et al. (2018)	EER	10.9	LSTM with acoustic embeddings
Mallidi et al. (2018)	EER	5.2	Fully-connected DNN of two LSTMs (acoustic embeddings, and 1-best embeddings) + ASR decoder features
Huang et al. (2019)	EER	-35.36% (rel.)	LSTM (ASR-decoder and *log-STFT256* features) + attention-Pooling
Tong et al. (2021)	EER	-41.1% (rel.)	ResLSTM (*log-STFT256* features) + attention aggregation
VACC			
Siegert and Krüger (2021)	UAR	81.97	SVM (*emobase*)
Akhtiamov et al. (2019)	UAR	90.10	SVM (450 Low-Level-Descriptors (LLDs) of *ComParE* with recursive feature elimination) and mixup data augmentation (VACC + SVC)
RBC			
Akhtiamov et al. (2020)	UAR	62.80	SVM metamodel classifier from linear SVM (1,600 *ComParE* features with recursive feature elimination), radial SVM (ASR decoder features), LSTM (128 *ComParE*-LLDs), and e2e
Baumann and Siegert (2020)	F1-score	65.50	biRNN with attention layer (segmental information, *MFCC*s, and *FFV*)
Siegert et al. (2021)	UAR	85.77	speaker-depended continuous learning framework (*emobase*)

The architecture used in Norouzian et al. (2019) differs significantly from the other studies reported in this survey and employs the capabilities of both CNN and LSTM architectures. The audio files are transformed into a visual representation of log-mel filterbanks from the original recording, which is directly interpreted by a CNN. As the information is framewise fed to this system, it needs to be aggregated for the full utterance. Therefore, a biLSTM network is utilized by the authors. Additionally, there exists an attention layer, which focuses on the important parts of the extracted information from the architecture to this point. The final classification is then conducted using a densely connected feed-forward network. This complex deep learning network achieved an EER of 16.25% on a real dataset it trained on and 15.62% on the dataset it was not trained on, which showed remarkable generalization effects.

In Mallidi et al. (2018) a fully connected feed-forward DNN is trained to combine the two individual LSTMs classifier capturing the acoustic embeddings and the ASR 1-best hypotheses. Additionally, features from the ASR decoder are fed into the DNN as well. With this approach, the authors achieve a final EER of 5.2%. The classifier architecture in Tong et al. (2021) consisted of an adapted ResNet, which also used a CNN for internal layers, but goes into deeper layered architectures by allowing the data from each layer to also jump over the convolutional step of each iteration. This allows for information to remain unchanged in the system for the latter computational steps. This approach was then fed into a three-layered LSTM network with an aggregation layer at the end. The final classification step is then again utilized by an independent DNN. This resulted in a declared relative improvement of 8.7% ACC compared to the original model, as well as a relative reduction of 41.1% EER, in absolute numbers, this results in an EER of 5.48%.

4.4.5. Performance Comparison

Although the reported performance among all studies cannot be compared directly, due to the usage of different validation schemes (leave-one-speaker-out vs. cross-validation) and different evaluation measures (EER, ACC, F1-score, UAR), a general trend in the performance can be drawn from the reported results of the identified studies. Therefore, in **Table 5** selected studies employing the same datasets (and the same evaluation measure), are depicted. This allows a comparison of promising approaches as well as a forecast for future trends. Especially, the outstanding results of Akhtiamov and Palkov (2018), Tong et al.

(2021) on existing data and in-house data, respectively, show that the prosodic addressee detection already achieves very good results. Noteworthy are also the results presented in Siegert et al. (2021), using a specialized dataset of identical human-human and human-device utterances with nearly similar prosody, as it can be expected for future voice assistants having natural interaction abilities (Biundo and Wendemuth, 2016).

The comparison of the performances across the different datasets further allows answering the question to which extent the number of features, the types of features, and the model architecture influence the performance. Especially for the earlier studies of Shriberg et al., it is apparent that the selection of proper feature types contributes considerably to the recognition performance (Shriberg et al., 2013). Although, larger feature sets are utilized in the latter studies, also the use of sophisticated model architectures contributes to the improvement of the performance. In comparison to Batliner et al. (2009), where 100 features and a Linear Discriminant Classifier is used, later studies on the same dataset using SVMs with 1,000 features archive an improvement of more than 10% absolute (e.g., Akhtiamov et al., 2017b; Pugachev et al., 2018). The influence of a sophisticated model architecture on the classification performance can be seen in Baumann and Siegert (2020) where a biRNN with attention layer outperforms a metamodel of linear models despite the fact that the metamodel used much more features (Akhtiamov et al., 2020). Additionally, the selection of specific feature types seems to have a substantial influence. Thus, it is still unclear whether optimized feature (selection) or sophisticated model architectures are more promising to further improve the performance and, as it can also be seen in the studies of this survey, both approaches are used in parallel.

4.4.6. Summary–The Evolution of Machine Learning Methods

Regarding the presented studies, it can be stated that the use of the different classification systems went parallel to the development of machine learning methods. The early studies were using varying classical methods, for example, Bayesian Networks (Takemae and Ozawa, 2006), Linear Discriminant Classifiers (Batliner et al., 2009; Hayakawa et al., 2016a), or Random Forests (Vinyals et al., 2012). Later on, the usage of SVM appeared as a baseline for many applications, beginning (Le Maitre and Chetouani, 2013). Furthermore, several studies employed an SVM as a meta-classifier (e.g., Akhtiamov et al., 2017c, 2020; Akhtiamov and Palkov, 2018). Recently, beginning with Akhtiamov et al. (2017c) many systems also employed LSTMs as a recurrent component in their classification process, often in addition to CNNs or combined several classifiers with a final densely/fully connected network (e.g., Norouzian et al., 2019; Wang et al., 2020; Tong et al., 2021).

5. FUTURE APPROACHES

5.1. Limitations

Given the enormous variations in the description of the topic of addressee detection using terms like "device-directed" or "talking to machines" in this review, some further keywords could get missed, although the keyword was searched in the title and the abstract. As the resulting studies were then screened by reading only the title and/or abstract, studies where the speech description was not the main focus, may have been missed if "speech" or related terms were not in the title or abstract.

5.2. Generalization

Even though various feature sets and models are used and have shown their eligibility for one specific dataset, it is yet unclear if they will generalize to a new speaker that varies in age, geography, socio-economic level, and recording setup. Especially regarding the latter issue, several studies analyzed the performance regarding prosodic-affective recognition across different acoustic conditions as well as acoustic compression (Siegert et al., 2016; Gottschalk et al., 2020; Siegert and Niebuhr, 2021).

It can be assumed that at least the models by Mallidi et al. (2018), Huang et al. (2019), and Tong et al. (2021) are more broadly positioned due to the great amount of data utilized, but as nearly nothing is known about this data this will remain speculative. Furthermore, it can at least be assumed that the users of these voice-controlled far-field devices are selective to specific types of users (McLean and Osei-Frimpong, 2019; Koenecke et al., 2020).

5.3. Multimodal Learning

Even though this review focuses on speech, many studies provided multimodal classifiers trained on prosodic-acoustic and lexical/semantic data. Where applicable, these different results were reported as well for comparison to unimodal models.

5.4. Multi-User Setting

Due to the setup of distinguishing human-directed speech from device-directed speech, most of the data comes from a multi-user setting, where the user of the voice assistant is talking to both the technical system and another human being. Unfortunately, not in all studies, the setups are as clear, like settings where the interaction toward the technical system and toward the interlocutor is happening independently at two different points in time is possible. Furthermore, not much is known about the human interlocutors in most of the data sets, thus it remains unclear if and how they could influence the speech behavior of the users.

6. CONCLUSION

The large number of similar search terms that were used to identify the relevant studies, together with the large variability in index terms used by the authors of the studies, shows that this research area is very complex and has been addressed from multiple directions. After removing duplicate records and excluding mismatching studies, a total of 23 studies were reviewed which report on automatic addressee detection from speech using acoustic information. Most of the identified studies are not older than 10 years, showing the actuality of this emerging topic. The survey depicted a continuous development of the methods in the area of machine learning applications in terms

of datasets, features as well as classifier architectures. These three categories were also used alongside this survey.

Regarding the utilized datasets, we concluded that the availability of high-quality data was a crucial issue in the beginning. Today, on the one hand, huge datasets are utilized at least as in-house data (e.g., Norouzian et al., 2019; Tong et al., 2021), but without a public availability. On the other hand, specialized datasets inside and outside laboratory settings were recorded and distributed (e.g., Bohus and Horvitz, 2009; Siegert et al., 2018). These datasets are widely used in different identified studies (e.g., Tsai et al., 2015b; Baumann and Siegert, 2020). One dataset, that was already recorded in 2006 and still used as a reference, and therefore can be seen as a baseline dataset, is the SVC (Batliner et al., 2006). Furthermore, more studies used multiple datasets either for data augmentation or cross-corpora experiments to improve generalization and a broader application. Afterwards, we identified the same trends in feature sets and for the classification architectures as for speech-based machine learning in general (Ververidis and Kotropoulos, 2006; Nassif et al., 2019). The first identified studies utilized hand-crafted, rather small feature sets and classical classification architectures (Batliner et al., 2009; Shriberg et al., 2012). While the latter studies rely on more standardized (widely-known) and larger feature sets, specifically the *ComParE* feature set from the Interspeech 2013 challenge (Schuller et al., 2013) in combination with SVM as a more sophisticated classifier (e.g., Akhtiamov et al., 2017b; Akhtiamov and Palkov, 2018). Some studies further utilize feature elimination methods to gain specialized features (e.g., Hayakawa et al., 2016a; Akhtiamov and Palkov, 2018) or apply newly developed feature descriptions, such as *FFV*s (Baumann and Siegert, 2020). Through the parallel increased use of deep learning networks, on the one hand, special features such as *LFBE* or *log-STFT256* were utilized in conjunction with CNNs (Mallidi et al., 2018) and LSTMs (Huang et al., 2019).

As an offshoot of the on and off-talk detection of the first voice-controlled system, it developed into an independent research idea of separating device and human-directed speech. With the current increasing use of assistant systems, there was also an increase in new systems capable of removing the wake-word-dependency and similar external activation signals, without leading to a system prone to false activations. Thereby, the ease of use of actual voice assistants can be further simplified. The advantage of purely acoustic addressee detection can then be exploited in particular in multi-user interactions. The assistant in such a setting is then also able to recognize who (from the group) addressed the assistant in which situation without actively listening to every word we speak in its proximity. This is especially helpful to transform the use of voice assistants from relatively trivial activities such as collecting information, listening to music, or sending messages or calls to cases where voice assistants conduct larger dialogs and support the user to a greater extent. For example, to help patients communicate with their caregivers (Shu, 2019) or with remote diagnosis (Futurist, 2021). Another area of application are tutoring systems in educational contexts (Callaghan et al., 2019; Winkler et al., 2019) or for language acquisition systems. Furthermore, acoustic address recognition can also contribute to increased privacy and trust in voice assistants by minimizing false activations (Wienrich et al., 2021; Kisser and Siegert, 2022).

What is apparent from the different identified studies is the lack of proper benchmark data which is publicly available and of sufficient size. This is needed to on the one hand further improve the classification performance and on the other hand to overcome the limitations of current research with respect to possible data bias (Cramer et al., 2018; Koenecke et al., 2020). Therefore, we encourage creating open data sets, if possible through competitions, as they have shown to be highly productive in other settings.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work has been partly sponsored by the project Intention-based Anticipatory Interactive Systems (IAIS) funded by the European Funds for Regional Development (EFRE) and by the Federal State of Saxony-Anhalt, Germany, under the grant number ZS/2017/10/88785.

ACKNOWLEDGMENTS

We acknowledge support for the article processing charge by the Open Access Publication Fund of Otto von Guericke University Magdeburg.

REFERENCES

Ahuja, K., Kong, A., Goel, M., and Harrison, C. (2020). "Direction-of-voice (dov) estimation for intuitive speech interaction with smart devices ecosystems," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, UIST '20* (New York, NY: Association for Computing Machinery), 1121–1131.

Akhtiamov, O., and Palkov, V. (2018). "Gaze, prosody and semantics: relevance of various multimodal signals to addressee detection in human-human-computer conversations," in *Speech and Computer*, eds A. Karpov, O. Jokisch, and R. Potapova (Cham: Springer International Publishing), 1–10.

Akhtiamov, O., Sidorov, M., Karpov, A., and Minker, W. (2017a). "Speech and text analysis for multimodal addressee detection in human-human-computer interaction," in *Proceedings of the INTERSPEECH'17*, Stockholm, 2521–2525.

Akhtiamov, O., Sidorov, M., Karpov, A. A., and Minker, W. (2017b). "Speech and text analysis for multimodal addressee detection in human-human-computer interaction," in *Proceedings Interspeech 2017*, Stockholm, 2521–2525.

Akhtiamov, O., Siegert, I., Karpov, A., and Minker, W. (2019). "Cross-corpus data augmentation for acoustic addressee detection," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue* (Stockholm: Association for Computational Linguistics), 274–283.

Akhtiamov, O., Siegert, I., Karpov, A., and Minker, W. (2020). Using complexity-identical human- and machine-directed utterances to investigate addressee detection for spoken dialogue systems. *Sensors* 20, 2740. doi: 10.3390/s20092740

Akhtiamov, O., Ubskii, D., Feldina, E., Pugachev, A., Karpov, A., and Minker, W. (2017c). "Are you addressing me? multimodal addressee detection in human-human-computer conversations," in *Speech and Computer*, eds A. Karpov, R. Potapova, and I. Mporas (Cham: Springer International Publishing), 152–161.

Baraldi, S., Bimbo, A. D., Landucci, L., and Torpei, N. (2009). *Encyclopedia of Database Systems, Chapter Natural Interaction* (Boston, MA: Springer US), 1880–1885.

Batliner, A., Hacker, C., and Nöth, E. (2006). *To Talk or not to Talk with a Computer: On-Talk vs. Off-Talk*. SFB/TR 8 Report No. 010-09/2006. University of Bremen, Bremen, Germany.

Batliner, A., Hacker, C., and Nöth, E. (2009). To talk or not to talk with a computer. *J. Multimodal User Interfaces* 2, 171. doi: 10.1007/s12193-009-0016-6

Baumann, T., and Siegert, I. (2020). "Prosodic addressee-detection: Ensuring privacy in always-on spoken dialog systems," in *Proceedings of the Conference on Mensch Und Computer, MuC '20* (New York, NY: Association for Computing Machinery), 195–198.

Benesty, J., Sondhi, M. M., and Huang, Y. A. (2008). *Springer Handbook of Speech Processing*. Berlin; Heidelberg: Springer.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: a survey. *Speech Commun*. 56, 85–100. doi: 10.1016/j.specom.2013.07.008

Biundo, S., and Wendemuth, A. (2016). Companion-technology for cognitive technical systems. KI-Künstliche Intelligenz. 30, 71–75. doi: 10.1007/s13218-015-0414-8

Bohus, D., and Horvitz, E. (2009). "Dialog in the open world: Platform and applications," in *Proceedings of the 2009 International Conference on Multimodal Interfaces, ICMI-MLMI '09*, Cambridge, MA, 31–38.

Bohus, D., and Horvitz, E. (2010). "Facilitating multiparty dialog with gaze, gesture, and speech," in *Proceedings of the 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI '10)* (New York, NY: Association for Computing Machinery).

Busso, C., Georgiou, P. G., and Narayanan, S. S. (2007). "Real-time monitoring of participants' interaction in a meeting using audio-visual sensors," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *Vol. 2* (Honolulu, HI: IEEE), II-685-II-688.

Callaghan, M. J., Bengloan, G., Ferrer, J., Cherel, L., El Mostadi, M. A., Gomez Eguíluz, A., et al. (2019). "Voice driven virtual assistant tutor in virtual reality for electronic engineering remote laboratories," in *Proceedings of the 15th International Conference on Remote Engineering and Virtual Instrumentation* (Düsseldorf: Springer), 570–580.

Casillas, M., Amatuni, A., Seidl, A., Soderstrom, M., Warlaumont, A. S., and Bergelson, E. (2017). "What do babies hear? analyses of child- and adult-directed speech," in *Proceedings of Interspeech 2017*, Stockholm, 2093–2097.

Casillas, M., and Frank, M. C. (2017). The development of children's ability to track and predict turn structure in conversation. *J. Mem. Lang.* 92, 234–253. doi: 10.1016/j.jml.2016.06.013

Chung, H., Iorga, M., Voas, J., and Lee, S. (2017). Alexa, can I trust you? *Computer* 50, 100–104. doi: 10.1109/MC.2017.3571053

Cramer, H., Garcia-Gathright, J., Springer, A., and Reddy, S. (2018). Assessing and addressing algorithmic bias in practice. *Interactions* 25, 58–63. doi: 10.1145/3278156

Dojchinovski, D., Ilievski, A., and Gusev, M. (2019). "Interactive home healthcare system with integrated voice assistant," in *Proceedings of the 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, 284–288.

Dubois, D. J., Kolcun, R., Mandalari, A. M., Paracha, M. T., Choffnes, D., and Haddadi, H. (2020). "When speakers are all ears: characterizing misactivations of iot smart speakers," in *Proceedings of the Privacy Enhancing Technologies Symposium (PETS)*, Montreal.

Durana, P., Perkins, N., and Valaskova, K. (2021). Artificial intelligence data-driven internet of things systems, real-time advanced analytics, and cyber-physical production networks in sustainable smart manufacturing. *Econ. Manag. Finan. Markets* 16, 20. doi: 10.22381/emfm16120212

Everts, E. (2004). *Discourse and Technology. Multimodal Discourse Analysis, Chapter Modalities of Turn-Taking in Blind/Sighted Interaction: Better to Be Seen and Not Heard*? Georgetown, TX: Georgetown University Press.

Eyben, F., Weninger, F., Wöllmer, M., and Schuller, B. (2013). *openSMILE-the Munich open Speech and Music Interpretation by Large Space Extraction toolkit. Number 2*. Munich: Technical University Munich.

Futurist, T. M. (2021). "The top 12 healthcare chatbots," in *Medical Future*. Available online at: https://medicalfuturist.com/top-12-health-chatbots.

Garvey, C., and Berninger, G. (1981). Timing and turn taking in children's conversations. *Discourse Process.* 4, 27–57. doi: 10.1080/01638538109544505

Gilmartin, E., Cowan, B., Vogel, C., and Campbell, N. (2018). Explorations in multiparty casual social talk and its relevance for social human machine dialogue. *J. Multimodal User Interfaces* 12, 297–308. doi: 10.1007/s12193-018-0274-2

Glodek, M., Honold, F., Geier, T., Krell, G., Nothdurft, F., Reuter, S., et al. (2015). Fusion paradigms in cognitive technical systems for human-computer interaction. *Neurocomputing* 161, 17–37. doi: 10.1016/j.neucom.2015.01.076

Gottschalk, M., Höbel, J., Siegert, I., Verhey, J. L., and Wendemuth, A. (2020). "Filtering-based analysis of spectral and temporal effects of room modes on low-level descriptors of emotionally coloured speech," in *Elektronische Sprachsignalverarbeitung 2020. Tagungsband der 31. Konferenz, volume 95 of Studientexte zur Sprachkommunikation* (Magdeburg: TUDpress), 219–226.

Gruzauskas, V., Krisciunas, A., Calneryt,e, D., Navickas, V., and Koišová, E. (2020). Development of a market trend evaluation system for policy making. *J. Competit.* 12, 22–37. doi: 10.7441/joc.2020.02.02

Haji, T., Horiguchi, S., Baer, T., and Gould, W. (1986). Frequency and amplitude perturbation analysis of electroglottograph during sustained phonation. *J. Acoust. Soc. Am.* 80, 58–62. doi: 10.1121/1.394083

Hayakawa, A., Haider, F., Luz, S., Cerrato, L., and Campbell, N. (2016a). Talking to a system and oneself: a study from a speech-to-speech, machine translation mediated map task. *Proc. Speech Prosody* 2016, 776–780. doi: 10.21437/SpeechProsody.2016-159

Hayakawa, A., Luz, S., Cerrato, L., and Campbell, N. (2016b). The "ILMT-s2s corpus — a multimodal interlingual map task corpus," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (Portorož: European Language Resources Association (ELRA)), 605–612.

Heck, L., Hakkani-Tür, D., Chinthakunta, M., Tur, G., Iyer, R., Parthasacarthy, P., et al. (2013). "Multimodal conversational search and browse," in *IEEE Workshop on Speech, Language and Audio in Multimedia*, Marseille.

Horcher, G. (2018). *Woman Says her Amazon Device Recorded Private Conversation, Sent it Out to Random Contact*. KIRO7. Available online at: https://perma.cc/8CG9-3M3G.

Huang, C.-W., Maas, R., Mallidi, S. H., and Hoffmeister, B. (2019). "A study for improving device-directed speech detection toward frictionless human-machine interaction," in *Proceedings of the INTERSPEECH'19*, Graz, 3342–3346.

Jovanovic, N., op den Akker, R., and Nijholt, A. (2006). "Addressee identification in face-to-face meetings," in *Proceedings of the 11th EACL*, Olomouc, 169–176.

Kinsella, B. (2020). *Nearly 90 Million u.s. Adults Have Smart Speakers, Adoption Now Exceeds One-Third of Consumers*. Sonderborg: voicebot.ai.

Kisser, L., and Siegert, I. (2022). "Erroneous reactions of voice assistants "in the wild" – first analyses," in *Elektronische Sprachsignalverarbeitung 2022. Tagungsband der 33. Konferenz, volume 103 of Studientexte zur Sprachkommunikation* (Sonderborg: TUDpress), 113–120.

Kleinberg, S. (2018). 5 ways voice assistance is shaping consumer behavior. *think with Google*.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., et al. (2020). Racial disparities in automated speech recognition. *Proc. Natl. Acad. Sci. U.S.A.* 117, 7684–7689. doi: 10.1073/pnas.1915768117

Kumar, D., Paccagnella, R., Murley, P., Hennenfent, E., Mason, J., Bates, A., et al. (2018). "Skill squatting attacks on Amazon Alexa," in *27th USENIX Security Symposium (USENIX Security 18)* (Baltimore, MD), 33–47.

Lăzăroiu, G., Kliestik, T., and Novak, A. (2021). Internet of things smart devices, industrial artificial intelligence, and real-time sensor networks in sustainable cyber-physical production systems. *J. Self Govern. Manag. Econ.* 6, 20. doi: 10.22381/jsme9120212

Lalanne, D., Nigay, L., Palanque, p., Robinson, P., Vanderdonckt, J., and Ladry, J.-F. (2009). "Fusion engines for multimodal input: a survey," in *Proceedings of the 2009 International Conference on Multimodal Interfaces, ICMI-MLMI '09* (New York, NY: Association for Computing Machinery), 153–160.

Le Maitre, J., and Chetouani, M. (2013). Self-talk discrimination in human-robot interaction situations for supporting social awareness. *Int. J. Soc. Rob.* 5, 277–289. doi: 10.1007/s12369-013-0179-x

Liptak, A. (2017). Amazon's Alexa started ordering people dollhouses after hearing its name on TV. *The Verge*.

Lunsford, R., and Oviatt, S. (2006). "Human perception of intended addressee during computer-assisted meetings," in *Proceedings of the 8th ACM ICMI* (Banff, AB), 20–27.

Mahajan, K., Large, D. R., Burnett, G., and Velaga, N. R. (2021). Exploring the benefits of conversing with a digital voice assistant during automated driving: a parametric duration model of takeover time. *Transport. Res. F Traffic Psychol. Behav.* 80, 104–126. doi: 10.1016/j.trf.2021.03.012

Malkin, N., Deatrick, J., Tong, A., Wijesekera, P., Egelman, S., and Wagner, D. (2019). Privacy attitudes of smart speaker users. *Privacy Enhancing Technol.* 2019, 250–271. doi: 10.2478/popets-2019-0068

Mallidi, S. H., Maas, R., Goehner, K., Rastrow, A., Matsoukas, S., and Hoffmeister, B. (2018). "Device-directed utterance detection," in *Proceedings of the INTERSPEECH'18*, 1225–1228, Hyderabad.

Martin, J. L., and Tang, K. (2020). "Understanding racial disparities in automatic speech recognition: the case of habitual "be"," in *Proceedings of Interspeech 2020*, Shanghai, 626–630.

Mayo, C., Aubanel, V., and Cooke, M. (2012). "Effect of prosodic changes on speech intelligibility," in *Proceedings of the INTERSPEECH'12*, Portland, OR, 1708–1711.

McLean, G., and Osei-Frimpong, K. (2019). Hey alexa... examine the variables influencing the use of artificial intelligent in-home voice assistants. *Comput. Hum. Behav.* 99, 28–37. doi: 10.1016/j.chb.2019.05.009

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. (2009). Moher d, liberati a, tetzlaff j, altman dg, group ppreferred reporting items for systematic reviews and meta-analyses: the prisma statement. *PLoS Med.* 6, e1000097. doi: 10.1371/journal.pmed.1000097

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., and Shaalan, K. (2019). Speech recognition using deep neural networks: a systematic review. *IEEE Access* 7, 19143–19165. doi: 10.1109/ACCESS.2019.2896880

Norouzian, A., Mazoure, B., Connolly, D., and Willett, D. (2019). "Exploring attention mechanism for acoustic-based classification of speech utterances into system-directed and non-system-directed," in *Proceedings of the IEEE ICASSP-2019*, Brighton, 7310–7314.

Olson, D. L., and Delen, D. (2008). *Advanced Data Mining Techniques*. Berlin; Heidelberg: Springer.

Oppermann, D., Schiel, F., Steininger, S., and Beringer, N. (2001). "Off-talk - a problem for human-machine-interaction," in *7th European Conference on Speech Communication and Technology, EUROSPEECH-2001, September 3-7, 2001* (Aalborg), 2197–2200.

Osborne, J. (2016, July 20). Why 100 million monthly cortana users on windows 10 is a big deal. *TechRadar*.

Ouchi, H., and Tsuboi, Y. (2016). "Addressee and response selection for multi-party conversation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, TX: ACL), 2133–2143.

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-factor to ROC, informedness, markedness and correlation. *arXiv*. cs.LG. doi: 10.48550/ARXIV.2010.16061

Pugachev, A., Akhtiamov, O., Karpov, A., and Minker, W. (2018). "Deep learning for acoustic addressee detection in spoken dialogue systems," in *Artificial Intelligence and Natural Language*, eds A. Filchenkov, L. Pivovarova, and J. Žižka (Cham: Springer International Publishing), 45–53.

Schönherr, L., Golla, M., Eisenhofer, T., Wiele, J., Kolossa, D., and Holz, T. (2020). Unacceptable, where is my privacy? Exploring Accidental Triggers of Smart Speakers. *arXiv:2008.00508*. cs.CR. doi: 10.48550/arXiv.2008.00508

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., et al. (2017). "The INTERSPEECH 2017 computational paralinguistics challenge: addressee, cold and snoring," in *Proceedings of the INTERSPEECH-2017* (Stockholm), 3442–3446.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, Autism," in *Proceedings of the INTERSPEECH-2013* (Lyon), 148–152.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Heck, L. (2012). "Learning when to listen: detecting system-addressed speech in human-human-computer dialog," in *Proceedings of the INTERSPEECH'12* (Portland, OR), 334–337.

Shriberg, E., Stolcke, A., and Ravuri, S. (2013). "Addressee detection for dialog systems using temporal and spectral dimensions of speaking style," in *Proceedings of the INTERSPEECH'13* (Lyon), 2559–2563.

Shu, C. (2019). "Cedars-Sinai puts Amazon Alexa in patient rooms as part of a pilot program," in *TechCrunch*. Available online at: https://techcrunch.com/2019/02/25/cedars-sinai-puts-amazon-alexa-in-patient-rooms-as-part-of-a-pilot-program/

Siegert, I. (2015). *Emotional and user-specific cues for improved analysis of naturalistic interactions* (Ph.D. thesis). Otto von Guericke University Magdeburg.

Siegert, I. (2021). "Effects of prosodic variations on accidental triggers of a commercial voice assistant," in *Proceedings of the INTERSPEECH'21*, Brno, 1674–1678.

Siegert, I., and Krüger, J. (2018). "How do we speak with ALEXA - subjective and objective assessments of changes in speaking style between HC and HH conversations," in *Kognitive Systeme*, Duisburg-Essen Publication Online.

Siegert, I., and Krüger, J. (2021). *Chapter "Speech Melody and Speech Content Didn't Fit Together"–Differences in Speech Behavior for Device Directed and Human Directed Interactions*. Cham: Springer International Publishing.

Siegert, I., Krüger, J., Egorow, O., Nietzold, J., Heinemann, R., and Lotz, A. (2018). "Voice assistant conversation corpus (VACC): a multi-scenario dataset for addressee detection in human-computer-interaction using Amazon's ALEXA," in *Proceedings of the 11th LREC* (Paris).

Siegert, I., Lotz, A., Maruschke, M., Jokisch, O., and Wendemuth, A. (2016). "Emotion intelligibility within codec-compressed and reduced bandwith speech," in *12. ITG-Fachtagung Sprachkommunikation*, Paderborn, 215–219.

Siegert, I., and Niebuhr, O. (2021). Case report: women, be aware that your vocal charisma can dwindle in remote meetings. *Front. Commun.* 5, 135. doi: 10.3389/fcomm.2020.611555

Siegert, I., Nietzold, J., Heinemann, R., and Wendemuth, A. (2019). "The Restaurant Booking Corpus - content-identical comparative human-human and human-computer simulated telephone conversations," in *Elektronische Sprachsignalverarbeitung 2019. Tagungsband der 30. Konferenz* (Dresden), 126–133.

Siegert, I., Weißkirchen, N., Krüger, J., Akhtiamov, O., and Wendemuth, A. (2021). Admitting the addressee detection faultiness of voice assistants to improve the activation performance using a continuous learning framework. *Cogn. Syst. Res.* 70, 65–79. doi: 10.1016/j.cogsys.2021.07.005

Siepmann, R., Batliner, A., and Oppermann, D. (2001). "Using prosodic features to characterize off-talk in human-computer interaction," in *Proceedings of the ISCA Tutorial and Research Workshop on Speech Recognition and Understanding, October 22-24, 2001* (Red Bank, NJ), 27.

Sinha, G., Shahi, R., and Shankar, M. (2010). "Human computer interaction," in *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, Goa, 1–4.

Sri Suvetha, C. S., Subiksha, S., and Suguna, M. (2022). "Automatic traffic sign detection system with voice assistant," in *2022 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, Coimbatore, 1-7.

Takemae, Y., Otsuka, K., and Mukawa, N. (2004). "An analysis of speakers' gaze behavior for automatic addressee identification in multiparty conversation and its application to video editing," in *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)* (Kurashiki: IEEE), 581–586.

Takemae, Y., and Ozawa, S. (2006). "Automatic addressee identification based on participants' head orientation and utterances for multiparty conversations," in *2006 IEEE International Conference on Multimedia and Expo* (Toronto, ON), 1285–1288.

Tong, X., Huang, C.-W., Mallidi, S. H., Joseph, S., Pareek, S., Chandak, C., et al. (2021). "Streaming ResLSTM with causal mean aggregation for

device-directed utterance detection," in *IEEE Spoken Language Technology Workshop* (Shenzhen).

Tsai, J., Stolcke, A., and Slaney, M. (2015a). "Multimodal addressee detection in multiparty dialogue systems," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (South Brisbane, QLD: IEEE), 2314–2318.

Tsai, J., Stolcke, A., and Slaney, M. (2015b). A study of multimodal addressee detection in human-human-computer interaction. *IEEE Trans. Multimedia* 17, 1550–1561. doi: 10.1109/TMM.2015.2454332

Vaidya, T., Zhang, Y., Sherr, M., and Shields, C. (2015). "Cocaine noodles: exploiting the gap between human and machine speech recognition," in *9th USENIX Workshop on Offensive Technologies (WOOT 15)* (Washington, DC).

Valaskova, K., Ward, P., and Svabova, L. (2021). Deep learning-assisted smart process planning, cognitive automation, and industrial big data analytics in sustainable cyber-physical production systems. *J. Self Govern. Manag. Econ*. 9, 9. doi: 10.22381/jsme9220211

Valli, A. (2007). *Notes on natural interaction*. Technical report, University of Florence, Italy.

van Turnhout, K., Terken, J., Bakx, I., and Eggen, B. (2005). "Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features," in *Proceedings of the 7th ACM ICMI*, Torento, 175–182.

Ververidis, D., and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Commun* 48, 1162–1181. doi: 10.1016/j.specom.2006.04.003

Vinyals, O., Bohus, D., and Caruana, R. (2012). "Learning speaker, addressee and overlap detection models from multimodal streams," in *Proceedings of the 14th ACM ICMI'12* (Santa Monica, CA), 417–424.

Wang, J., Kumar, R., Rodehorst, M., Kulis, B., and Vitaladevuni, S. N. P. (2020). "An audio-based wakeword-independent verification system," in *Proceedings of the INTERSPEECH'20*, Shanghai, 1952–1956.

Wienrich, C., Reitelbach, C., and Carolus, A. (2021). The trustworthiness of voice assistants in the context of healthcare investigating the effect of perceived expertise on the trustworthiness of voice assistants, providers, data receivers, and automatic speech recognition. *Front. Comput. Sci*. 3, 685250. doi: 10.3389/fcomp.2021.685250

Winkler, R., Bittner, E., and Söllner, M. (2019). "Alexa, can you help me solve that problem?-understanding the value of smart personal assistants as tutors for complex problem tasks," in *14. Internationale Tagung Wirtschaftsinformatik (WI 2019)* (Siegen: Universi- Universitätsverlag Siegen), 371–376.

Wu, M., Panchapagesan, S., Sun, M., Gu, J., Thomas, R., Vitaladevuni, S. N. P., et al. (2018). "Monophone-based background modeling for two-stage on-device wake word detection," in *Proceedings of the IEEE ICASSP-2018* (Calgary, AB: IEEE).

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). "mixup: beyond empirical risk minimization," in *Proceedings of International Conference on Learning Representations (ICLR)*, Vancouver, BC.

Zhang, N., Mi, X., Feng, X., Wang, X., Tian, Y., and Qian, F. (2019). "Dangerous skills: understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems," in *IEEE Symposium on Security and Privacy* (San Francisco, CA: IEEE), 1381–1396.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.