The Audio Interaction Technique and Its Applications In Small-Size Electrical Appliances

Yiwei Cao *

Gemdale Future School, Nanjing, 210000, China, 2025

* Corresponding Author Email: caoyiwei0811@outlook.com

Abstract. Audio interaction technology, which facilitates command of digital products through natural speech, serves as a cornerstone of modern intelligent systems, significantly enhancing daily life efficiency and user satisfaction. This paper details the foundational principles underpinning this technology, outlining the integrated three-stage process: Automatic Speech Recognition (ASR) decodes audio input into text, Natural Language Processing (NLP) analyzes this text to discern user intent and generate commands, and Text-to-Speech (TTS) synthesizes audible, human-like responses. The document then explores its transformative application and influence within the smart home ecosystem, providing specific analysis of its implementation in small-size appliances such as digital watches and smart speakers. These case studies demonstrate tangible benefits, including unparalleled hands-free convenience, enhanced safety, and improved accessibility for users with visual or mobility impairments. Finally, the paper discusses prospective future development directions, forecasting the mainstream adoption of low-power-consuming products and the continued expansion into new application scenarios, further optimizing and personalizing the domestic living experience.

Keywords: Audio Interaction Technology, ASR, NLP, TTS, Smart Home Applications.

1. Introduction

Audio interaction technology is the foundation of modern techniques, enabling human to command digital products and thus achieve a smarter lifestyle. For instance, the latest automatic driving cars can receive drivers' orders to navigate to destinations; in home appliances, a single lamp can be controlled precisely by users' voice commands. These cases have significantly increased the daily life efficiency and user satisfaction

The origins of audio interaction technology can be traced back to 1952, when Bell Labs developed Audrey, a system capable of recognizing Arabic numerals [1]. Later on, in 1962, International Business Machines Corporation (IBM) invented the Shoebox, the first machine that could perform simple mathematical calculations using voice [2]. The second generation of audio recognition system emerged in 1980s; this model significantly increased the accuracy and adaptation ability to different languages and audio features, thanks to the application of hidden Markov models (HMMs) and machine learning algorithms [3]. In modern days, the develop of artificial intelligence (AI) has largely enhanced the ability of audio interaction products. AI has enhanced this technique with its high-level ability of understanding, translation capabilities and developed network, making the product more human-like. In 1997, IBM launched its first speech recognition product, Via Voice. In 2011, Apple first added the intelligent voice assistant Siri to the iPhone 4s. At this point, smart voice technology became deeply integrated with mobile phones.

However, the audio interaction is not limited to phones. In 2014, Amazon launched Amazon Echo, the first smart speaker on the market, bringing intelligent voice technology into the home and fundamentally shifting how users interact with hardware [4]. Its voice interaction function is implemented through the built-in AI assistant Alexa, which supports various functions including music playback, real-time information queries, and smart home control. This is the first time an audio interaction product was launched into market, and it apparently guide the trend for a long time. Many companies soon joined the competition and quickly show out their own products, such as Google Home and Apple HomePod [5]. Another recent example is XIAODU, an AI home assistant that also

functions as a wireless speaker. Loaded with DuerOS AI system, XIAODU can recognize different languages and respond more accurately. When users call its name and state their needs, XIAODU can identify keywords among daily speech. It can also connect to other home appliances, such as lights, televisions and curtains. This provides dwellers a chance to control their houses with an order spoken, enhancing users' experience with a more efficient and vivid lifestyle. It is also particularly convenient for people with reduced mobility. These are all examples of this technique used in home appliances. A key future direction involves the development of low-power-consuming products capable of on-device processing, which is crucial for privacy, latency, and widespread adoption in smart homes beyond always-listening speakers [6].

In this paper, the principles of audio interaction technology will be explained firstly, including how detect human voice and convert it into what machines could understand. Then, introduce the application and influence of audio interaction technology in the field of smart home life. Finally, Finally, discuss the future development directions of audio interaction technology and more potential application scenarios in the field of smart home.

2. Principle of Audio Interaction

Audio interaction refers to the transmission of information between humans and devices through natural speech. A complete voice interaction requires going through ASR、NLP and TTS processes. Specifically, the ASR system decodes the input audio into text. Next, the NLP component converts the decoded audio into commands to let devices to understand. Finally, the TTS module outputs texts and voices to response users.

2.1. ASR

ASR is the first step to make the whole audio interaction system work, aiming to convert human speech signals into corresponding text. It is also the key to make the whole audio interaction system works. Existing studies generally believe that speech recognition is divided into two processes. The process of the ASR system is illustrated in Figure 1.

The initial stage involves pre-processing the input speech to enhance audio quality. This includes applying noise reduction techniques to minimize background interference, detecting and handling silent segments within the recording, and utilizing pre-emphasis filtering to amplify higher frequencies, thereby increasing overall clarity prior to further analysis.

After pre-processing, the audio signal is segmented into small, overlapping frames, typically lasting around 20 milliseconds in duration. From each frame, crucial acoustic characteristics are extracted to form a distinctive representation or fingerprint of the sound. This is commonly achieved by extracting features such as Mel Frequency Cepstral Coefficients or filter bank energies, which effectively capture the spectral properties essential for recognition.

The core recognition process employs deep learning-based acoustic models, which have dramatically improved recognition accuracy over previous methods [7]. such as models utilizing Connectionist Temporal Classification, Recurrent Neural Network Transducer, or Transformer architectures. These model maps the extracted sound features to sequences of phonemes, the fundamental units of pronunciation. To generate coherent and contextually appropriate transcriptions, a language model predicts plausible word sequences by leveraging contextual information. Finally, a decoder integrates outputs from both the acoustic and language models, employing search algorithms to determine the most probable sequence of words that corresponds to the spoken input.

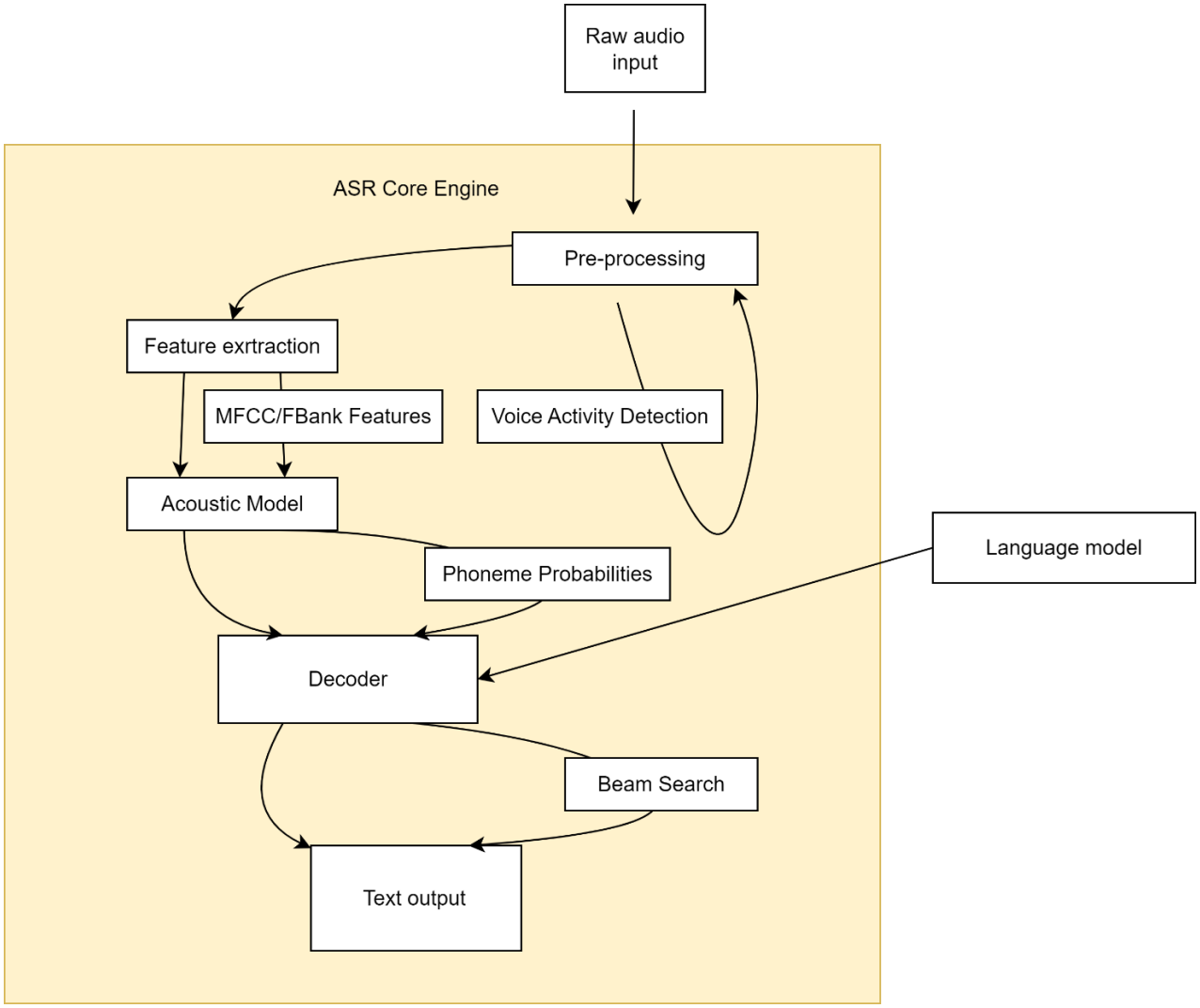


Fig. 1 The process of ASR (Picture credit : Original)

Overall, the language model and the acoustic model are the keys of the ASR process, guiding and determining the accuracy of the audio output. Additionally, the language model can also eliminate the problem induced by polyphonic words and homophones, minimize the inaccuracy.

2.2. NLP

Once the audio is decoded by ASR, the NLP system takes over to analyze, interpret, and generate responses based on verbal input instead of written text alone. This step serves as a bridge between speech and language understanding.

NLP is a critical technique of AI section. Human language is consisting of abstract information, while machines can only process coded data. Therefore, human language needs to be numerically converted. Fortunately, with help of developing AI, this process is getting more accurate.

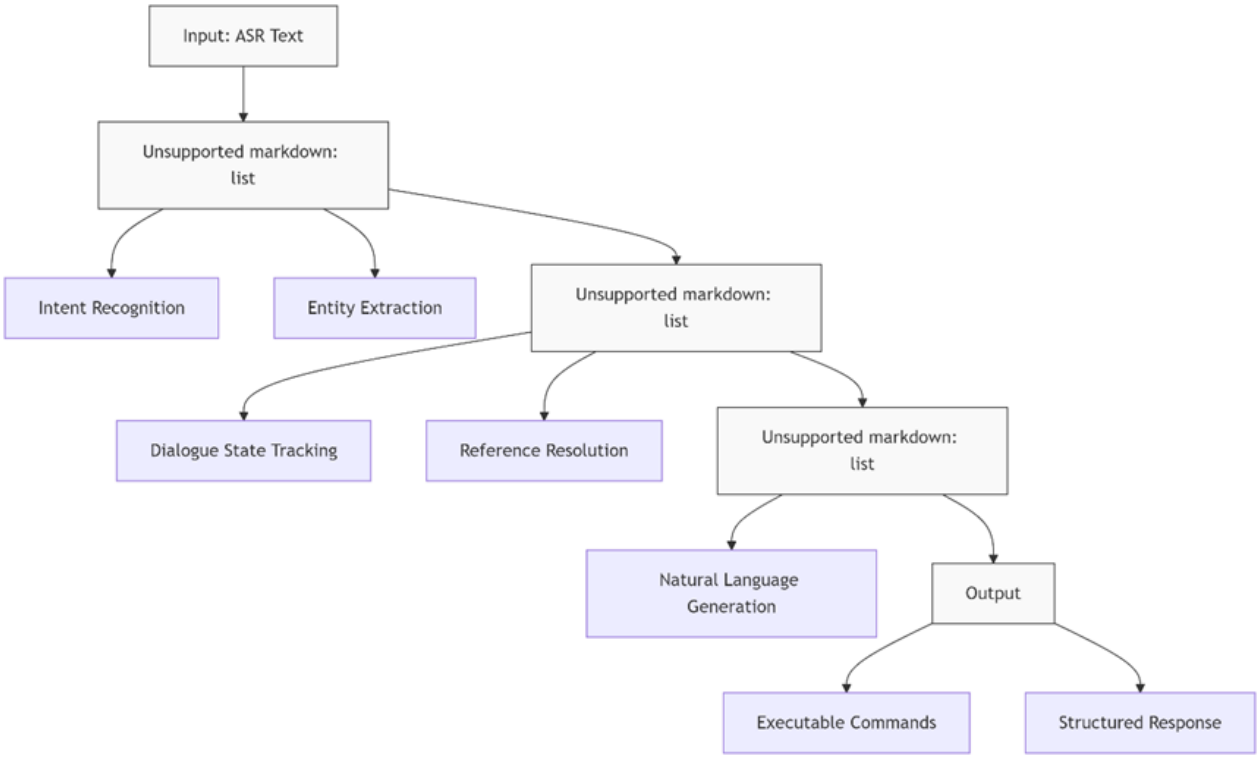


Fig. 2 The principle of NLP (Picture credit : Original)

The principle of NLP system is basically about the understanding of human languages. As shown in Figure 2, its main workflow proceeds as follows:

First, text data is collected from various sources and stored systematically. This raw text then undergoes cleaning and preprocessing: it is segmented into words or sentences, standardized to lowercase, stripped of common words and punctuation, and simplified to root word forms. Spelling corrections and abbreviation expansions are also applied.

Next, the cleaned text is converted into numerical formats computers can process. Methods include counting word frequencies, calculating term importance across documents, and creating vector representations that capture semantic meaning and relationships between words. Additional features identify sequences of words, grammatical structures, and contextual meanings.

Suitable algorithms are then chosen. Supervised models use labeled examples; unsupervised methods find hidden patterns; and pre-trained models like BERT apply existing knowledge to new tasks through fine-tuning, significantly advancing the field of natural language understanding [8]. These trained models then perform tasks like classifying text, identifying key entities, translating languages, or answering questions.

Finally, performance is evaluated using standard metrics. The system is refined through parameter adjustments, error analysis, and repeated improvements to ensures its reliability in real-world application.

By the way, AI models are crucial assistants in understanding human language. With machine learning algorithms. With developing of AI models, the accuracy of NLP translation can be improved significantly.

2.3. TTS

NLP prepares text for synthesis by determining pronunciation, sentence structure, and intended emphasis. TTS systems then transform this analyzed linguistic information into audible speech, using computational models to accurately reproduce human-like rhythm, intonation, and vocal characteristics.

This technology performs such conversion through a multi-stage computational pipeline. The core principle basically involves the following steps: raw text 、 extraction of linguistic features 、 generation of acoustic parameters and synthesis of human-like speech waveforms.

The process begins by preparing the input text. This involves expanding abbreviations, converting numbers to words, and segmenting the text into its smallest sound units. It also predicts the speech's rhythm, pitch patterns, and natural pauses.

Next, the processed linguistic information is transformed into detailed sound frequency patterns that represent speech energy over time. Deep learning models, such as WaveNet, generate these acoustic representations based on the text's sound units, creating highly natural-sounding speech [9]. These frequency patterns are then converted into audible speech. Specialized models synthesize the raw audio waveforms, effectively simulating the physical mechanisms of human vocal production to create natural-sounding speech.

This technology combines deep learning approaches to balance speed, reliability, and accuracy in creating synthetic speech. It enables practical applications like voice assistants and other spoken-response systems.

3. Applications on Small-Size Electrical Appliances

Audio interaction allows users to control electrical devices through voice commands, instead of physical touch or clicks of buttons. In modern time, more companies dived in to develop more products integrated with audio interaction, such as Apple in their Apple watches, and XIAOMI in their speakers. These products enhance users' daily convenience, with helps of internal setup AI assistants and consist process of audio transformations. The following section will elaborate on how this technology is applied to electrical appliances, especially on digital watches and speakers.

3.1. Digital Watches

In September 2013, Samsung published the Galaxy Gear, marking the first smartwatch to bring voice interaction as one of its core features to the mainstream market. In April 2015, the first apple watch was available for external sale. With the built-in AI assistant Siri, multiple commands can be applied to this watch, including setting alarm clocks and accessing navigations. During route navigation, a subtle notification can be issued when turning left or right, so users do not have to look at the screen when looking for directions.

As shown in Figure 3, the principle of audio interaction on digital watches centers on creating a bidirectional auditory channel. When a user speak out a command, The ASR system recognizes the voice input and begins processing, converting spoken words into actionable commands or queries interpretable by the device's operating system. Output is delivered via a miniature speaker or transducer, generating sounds ranging from simple alerts to synthesized speech. This synthesized speech is generated by TTS systems, converting textual information into audible speech. Additionally, the system employs signal processing techniques to manage ambient noise and optimize both audio capture and playback within the watch's compact form factor.

The application of audio interaction offers significant benefits for users. First and foremost, it substantially enhances accessibility, serving as a vital alternative interface for users with visual impairments or limited dexterity, a benefit noted in studies of voice assistant usage [10]. Furthermore, it delivers exceptional convenience and safety in situations where visual attention is occupied or hands are occupied, such as while driving, cooking, or exercising. In these situations, users can check information like the time, notifications, or fitness metrics, or issue commands such as setting a timer, all without needing to glance at or interact with the watch physically. This hands-free, eyes-free modality allows users to maintain focus on their primary task while efficiently interacting with their device through voice input and auditory feedback.

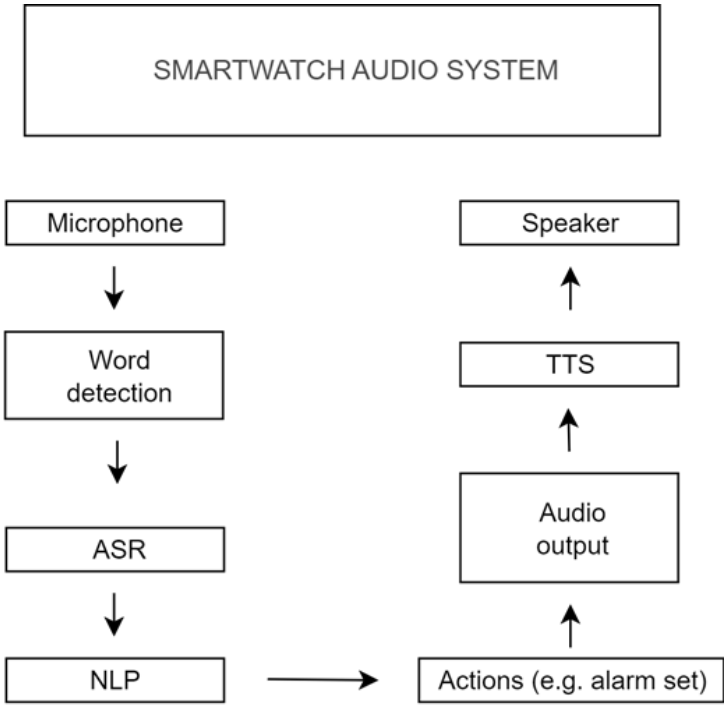


Fig. 3 The audio interaction works flow of a smart watch (Picture credit : Original)

3.2. Digital Speakers

Audio interaction in digital speakers, commonly experienced through smart speakers like Amazon Echo or Google Nest, involves a sophisticated process that transforms spoken commands into actionable outcomes. This technique, primarily driven by voice user interfaces (VUIs), relies on a seamless pipeline of hardware and software working in concert.

The process of a digital speaker is illustrated in Figure 4. It begins with far-field microphone arrays. These are not simple microphones but a set of strategically placed mics designed to overcome the challenges of home environments. Employing techniques like beamforming, they can pinpoint the direction of a user's voice, while acoustic echo cancellation filters out the music or audio the speaker itself is playing. This enables the device to clearly detect the "wake word" (e.g., "Xiaodu") even amid noise, allowing it to detect the voices when users call its name and reply accordingly.

Once the wake word is detected, the microphone array actively captures the subsequent voice command. This audio snippet is digitized, and for more complex processing tasks, it is compressed and securely transmitted to powerful cloud servers. Such offloading is crucial, as the speech recognition and natural language understanding (NLU) required are computationally intensive.

In the cloud, ASR software converts the analog waveform of the user's speech into digital text. Next, NLP engines parse this text to discern intent and extract key information, identifying the core action and the entities involved. The system then executes the command by interfacing with the appropriate service—whether a music streaming application programming interface, a smart home protocol, or its own internal knowledge graph for queries. Finally, the technique often completes the interaction with an audio response. This response is generated either via TTS engines that create a synthetic but natural-sounding voice to confirm an action or answer a question, or by playing a concise pre-recorded audio cue.

In essence, audio interaction transforms the speaker from a passive output device into an intelligent conversational agent. It operates as a continuous cycle of listening, processing, acting, and providing audible feedback, creating a hands-free, intuitive way to interact with technology.

Over time, the functions of a digital speaker are no longer restricted on playing music only. Thanks to their built-in neural network systems, users can bond them with housewares--such as televisions, curtains and doors--enabling voice control: a single spoken command allows these devices to operate autonomously.

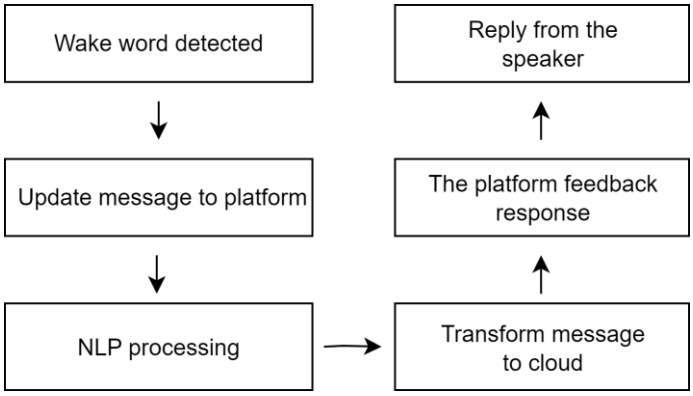


Fig. 4 The process of a digital speaker (Picture credit : Original)

4. Conclusion

For Conclusion Audio interaction technology, which enables voice control of digital devices, has evolved significantly since its inception in the mid-20th century. Its functionality relies on a core technical pipeline: ASR decodes audio into text, NLP derives meaning and intent from that text, and TTS generates audible responses. This process allows machines to understand and execute human vocal commands effectively.

The impact of this technology is profoundly evident in the smart home sector, particularly in small appliances like digital watches and smart speakers. These implementations offer enhanced user convenience, accessibility, and hands-free operation, transforming how people interact with their domestic environment. The future of audio interaction points towards its continued evolution, with low-power devices and expanded integration promising to further optimize daily life.

References

[1] Davis K. H., Biddulph R., Balashek S., et al. Automatic Recognition of Spoken Digits. The Journal of the Acoustical Society of America, 1952, 24 (6): 637–642.

[2] IBM Archives. IBM Shoebox. IBM Corporate Archives, 2023.

[3] Juang B. H., Rabiner L. R. Automatic Speech Recognition–A Brief History of the Technology Development. Elsevier Encyclopedia of Language and Linguistics, 2005.

[4] Hoy M. B., Alexa, Siri, Cortana, et al. An Introduction to Voice Assistants. Medical Reference Services Quarterly, 2018, 37 (1): 81–88.

[5] Huang W. B., Zhang M. Z., Sun J. Y., et al. Exploration of Application Security for Voice Recognition and Interaction Technology in Smart Home Appliances. Electric Appliances, 2021

[6] Xu M., Qian M., Sun Q., et al. Towards Energy Efficient Speech Recognition on Smartphones. 2020 IEEE/ACM Symposium on Edge Computing (SEC), 2020.

[7] Hinton G., Deng L., Yu D., et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine, 2012, 29 (6): 82–97.

[8] Devlin J., Chang M. W., Lee K., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv, 2018:1810: 04805,

[9] van den Oord A., Dieleman S., Zen H., et al. WaveNet: A Generative Model for Raw Audio. arXiv preprint arXiv, 2016, 1609: 03499,

[10] Pradhan A., Mehta K., Findlater L. "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018.