



“Foggy sounds like nothing” — enriching the experience of voice assistants with sonic overlays

Margarita Esau-Held¹ · Andrew Marsh¹ · Veronika Krauß¹ · Gunnar Stevens¹

Received: 31 March 2022 / Accepted: 19 March 2023 / Published online: 6 June 2023
© The Author(s) 2023

Abstract

Although Voice Assistants are ubiquitously available for some years now, the interaction is still monotonous and utilitarian. Sound design offers conceptual and methodological research to design auditive interfaces. Our work aims to complement and supplement voice interaction with *sonic overlays* to enrich the user experience. Therefore, we followed a user-centered design process to develop a sound library for weather forecasts based on empirical results from a user survey of associative mapping. After analyzing the data, we created audio clips for seven weather conditions and evaluated the perceived combination of sound and speech with 15 participants in an interview study. Our findings show that supplementing speech with soundscapes is a promising concept that communicates information and induces emotions with a positive affect for the user experience of Voice Assistants. Besides a novel design approach and a collection of sound overlays, we provide four design implications to support voice interaction designers.

Keywords Voice assistants · Sound design · User experience · Sonification · User study · Empirical design

1 Introduction

For several years now, households talk to Voice Assistants (VAs) in their homes and welcomed them as everyday companions [1–4]. Usually, most users use them predominantly to control and access home appliances and internet-based services [1, 5, 6], e.g., playing music, setting alarms, requesting weather forecasts, or asking for specific information [5]. By now, VAs have a significant contribution to the consumption of and interaction with information [1].

The progress in speech synthesis [7–10] and voice design [11] allows to make voices more human-like [11], less annoying [12], more appealing [13], more charismatic [14], or provide contextual cues implicitly [11]. In addition, the new opportunities offer designers to play with gender stereotypes [15], enable voice branding [16], or enrich the voice experience in general [17].

However, most users expect efficient and convenient interaction in a utilitarian sense as past experiences have disappointed them due to a lack of personal bonding and emotions [18]. Apart from considering the voice interaction as boring and monotone [19], users hope for a lively assistant, resembling a friend, that can express opinions and emotions itself as well as engage in a conversation [18].

In addition, the auditive channel bears potential in making use of sound design. Several researchers propose to explore interaction and experience beyond the dichotomie of human and machine and establish new design approaches for voice interaction [4]. Meanwhile, further researchers emphasize to integrate more sound design as well [20, 21]. The principles of sound design as there are sonification of data and interactions [22], musical expressions [23], the design of earcons and auditory icons [24, 25] represent great potential to enrich and enhance the current state of VAs. As stressed by Fagerlöhn and Liljedahl: “Sound design can be described

Margarita Esau-Held, Andrew Marsh, Veronika Krauß and Gunnar Stevens contributed equally to this work.

✉ Margarita Esau-Held
margarita.esau@uni-siegen.de

Andrew Marsh
andrew.marsh@student.uni-siegen.de

Veronika Krauß
veronika.krauss@uni-siegen.de

Gunnar Stevens
gunnar.stevens@uni-siegen.de

¹ Verbraucherinformatik Research Group, University of Siegen, Kohlbettstr. 15, Siegen 57072, Nord-Rhine Westphalia, Germany

as an inherently complex task, demanding the designer to understand, master and balance technology, human perception, aesthetics and semiotics.” [26].

While sound and the sonification of data could supplement the repertoire of speech synthesis and voice design by communicating information and expressing [22, 23], e.g., moods, atmospheres, emotions, interaction designers have not systematically adopted these extra options, so far. In this light, we draw from concepts and theories of sound design to explore our following two research questions:

RQ1 How might sound add to the user experience of Voice Assistants?

RQ2 How can we use sonification of data in information design for Voice Assistants?

In our work, however, we consider sound in its serving function to illustrate and enrich what is spoken by a Voice Assistant. In other words, we focus on the overlay quality of sound as a supplement to the speech output. As weather forecasts are a frequently used service of VAs, we decided to investigate this use case and its sonification. Therefore, we first conducted a user survey with 33 participants to empirically gather associative concepts and sounds for seven perceptible weather conditions. In the next step, we analyzed the design material and developed a sound library of seven distinct audio clips that illustrate our concept of sonic overlays. Finally, we evaluated and discussed our library with 15 participants in a qualitative interview study.

Our work shows that complementing voice interaction with illustrative soundscapes enriches the communication of VAs and is appreciated by potential users. As our empirical findings reveal, layering sound and speech needs special consideration of the relation of both and in light of the intended message. Therefore, we propose a user-centered design approach grounded in sound design that employs conceptual associations and the combination of iconic, abstract, and symbolic sounds. Sound Overlays, as outlined in this paper, could be used as an alternative to the advancements in speech science that focus on the modulation of emotions through the use of voice and speech as a design material. Furthermore, implementing a sound design in voice interaction might complement the emotional tone of voice of VAs in future designs. Soundscapes in voice interaction design add to the atmosphere of speakers to tell thrilling stories, as we know from sound design practices of modern media. Finally, we propose four design implications: Investigating soundscapes for voice interaction design (1), supplementing vocal messages by sound (2), aiming for authentic soundscapes (3), and finding a balance between expressiveness and informativeness as well as coping with trade-offs between clarity and sonification of information (4).

2 Related work

Our work is grounded in the following research fields in particular: VAs and Voice Interaction Design (see Section 2.1), earcons and sonic information design (see Section 2.2), and the design of sound effects and for sonic experiences in general (see Section 2.3). The first field focuses especially on the use of speech to enable natural conversational interaction with the user and addresses advancements in speech sciences to reflect on vocal speech as a key design material in voice interaction design. The second field deals with the auditory sense as an additional channel to encode and convey information. In terms of this work, we understand *encoding* of information as the process of using auditory channels to express information that humans can process with their auditory senses and understand in a meaningful way. Contrasting to the previous perspectives, the latter focuses on the effect and use of soundscapes in related fields of HCI and investigates the use of sound effects to enrich the experience of interactive media. To the best of our knowledge, only a few studies adopt concepts from sound design in the context of voice assistance and voice interaction design. In particular, current voice interaction research focuses on speech exclusively to make the voice output more natural and informative.

2.1 Voice interaction design

Voice interaction design represents a new type of interaction [19] that is primarily concerned with encoding and conveying information in spoken language. Particularly, the text-to-speech capabilities of current Natural Language Processing (NLP) machines [27, 28] enable and drive this emergence and growth of voice-first applications. The ephemeral character of speech-embodied information in comparison to text reveals different challenges of information communication by VAs, such as cognitive load or dead end conversations [2, 4, 20]. Due to a lacking persistent manifestation, cognitive load is increased and listeners are required to deeply focus in order to process and react to information [20]. Grice [29] argues that communication practices should always consider the quantity (right amount) and quality (speaking the truth) of information, as well as sharing only relevant information with a maximum of clarity.

However, user expectations regarding the capabilities of VAs remain frequently unfulfilled and cause disappointment and frustration as they expect an effortless and engaging exchange of information [18, 30]. Often, well-known usability issues like limited NLP and speech recognition, system errors, misunderstandings, and failed feedback cause this phenomenon [6, 31, 32]. As a result, this leads to an interaction style that is based on “guessing and exploration [rather] than knowledge recall or visual aids” [31]. Additionally, this type of conversational interaction does not feel

natural, and lacks sufficient positive experiences to motivate users to engage frequently [18]. Consequently, VAs need reliable usability to prevent users from negative experience [4, 31, 32], and furthermore research to investigate the positive aspects of user experience, which might contribute to an enchanting, playful, meaningful, and engaging interaction [33].

Accordingly, current research studies anthropomorphic effects and how to mimic human-human conversation successfully [32, 34], even though some research points to negative effects of too much human likeness [32]. Further experience dimensions for conversational agents might build on a more flexible attitude regarding the categories of “human” and “machine” [4, 35] and [3, 6, 19] should “fit into and around conversations” [19], and respectively routines of the users. We should understand speech as an act of performance, a kind of storytelling [34], and affective communication strategies [36] to enrich the interaction and stimulate experiences. For instance, new modes of articulation like “whispering” already extend the dimension of sonic experiences and prevent the VA from being perceived as boring and monotone [19].

Moreover, human information processing is not linear but complex. The Elaboration Likelihood Model [37, 38], for instance, stresses that humans process information via two routes: via the central route, people decode the content of the message by listening carefully to the semantics, the strength of arguments, and the credibility of included facts. In contrast, via the peripheral route, people respond emotionally to the message, where they are more likely to rely on general impressions, peripheral cues, and subliminal tones.

Affective and emotional speech research [14, 39], especially speech emotion recognition [7–9], emotional speech synthesis [10] and emotional speech production [40] represent an emerging research area addressing these subtle but vital aspects of communication. A body of work studies, for instance, how our voice and our way of speaking express a range of emotions like sadness, joy, anger, dearth, surprise, and boredom [10, 11, 41]. Furthermore, various studies have shown that speech and voice impact credibility, trust, charisma, attractiveness, likeability, and personality perception in general [11, 14, 42].

Research, machine learning in particular, also underlines the features responsible for communicating emotions. For example, research on emotional speech uncovered that acoustic levels such as frequency, bandwidth, pitching, intensity, loudness, speech rate, pausing, duration, and intonation of phonemes, words, and utterances influence the perception of emotions [7, 9, 39, 43]. Further, several linguistic and paralinguistic, among other more abstract features like gender, age, or accent, influence users’ perception of speech and voices [7, 11].

Regarding speech emotion design, researchers have specified various notation systems, such as the emotional markup language [44, 45], which allows designers to annotate parts of sentences to be spoken with a particular emotion. To support designers, Shi et al. [46] outline the concept of state-emotion mapping that may serve to drive human-VA conversational interaction. However, to save designers this additional annotation work, the researchers proposed a text-based emotion detection algorithm to contextually determine the emotional phrasing and pronunciation of sentences [39].

Our approach aims to supplement advances in speech science that focus on modulating emotions through speech to create engaging experiences between users and VAs by investigating alternative interaction design approaches.

2.2 Sound design and data sonification

Even though sound design is an active research field in the HCI community, there is a call for more scientific approaches to enable reproducible results [26]. So far, this field moves between craftsmanship and art and depends on skillful sound designers, as “Sound design can be described as an inherently complex task, demanding the designer to understand, master and balance technology, human perception, aesthetics and semiotics.” [26]. Sound is an integral part of media and system design to convey a captivating narrative, and an integral component for audiovisual storytelling [47].

Therefore data sonification represents an integral process to encode data and interactions so that the intended meaning is not misunderstood. According to Enge [48], sonification can be seen as “the use of nonspeech audio to convey information” [49], whereas visualization is understood as “the use of computer-supported, interactive, visual representations of abstract data to amplify cognition” [50]. Visualizations support a clear understanding of information, while sonification frequently allows for more interpretation despite its means to convey information [22]. Therefore, the most common approaches to auditorily encode information in interaction design are auditory icons and earcons [24]. A fundamental difference between auditory icons and earcons is that earcons can be considered to be arbitrary symbolic representations, while auditory icons can be regarded as analogical representations. Blattner et al. [24] defined earcons as “non-verbal audio messages used in the user-computer-interface to provide information to the user about some computer object, operation, or interaction”. Brewster further specifies that earcons are “abstract, synthetic tones that can be used in structured combinations to create auditory messages” [51].

The sonification of data is not only able to encode information but is also capable of expressing and inducing emotions. Depending on the design goal, inaccuracies may exist, as humans evaluate emotions very subjectively [22, 23]. Thereby, experiences are based on the affective and

functional perception of the design. This poses a challenge to research since it aims to investigate sonic elements and their impact objectively but competes with the narrative qualities of music and its affective and emotional impact [52]. While an interesting and positive experiential design may stimulate emotions, there will be a trade-off between the sonic experience and the clarity of the information [22]. The expression of emotions is defined by its psychophysical relationships between musical elements and perceptual impressions of the user. Further, capturing emotional expression in music is possible by focusing on a listener's agreement as no one can effectively deny their experience [23, 53, 54]. In contrast to expression, communication further depends on accurately recognizing the intended information and emotion [23, 55]. Therefore, our work aims to explore the relation between a clear understanding of information and the enrichment of emotions by combining sound and speech.

2.3 The role of sound design in modern immersive media

Following Simpson [20] and Sanchez-Chavez et al. [21], scholars argue that advanced methodologies and design principles for Conversational User Interfaces (CUI), e.g., interfaces for VAs, chatbots, are needed. So far, current designs follow engrained and trusted GUI principles to present and represent information without considering the dimensions of auditive information processing, for example, the ephemeral state of speech, memory, imagination, user interpretation [4, 20, 21]. Sanchez et al. [21] propose to even go beyond current conversational design “to include more nonverbal and paralinguistic elements” that could expand the design space further when considering sound interaction as a primary form of interaction.

In the light of the above, in most cases, sound is regarded as a complementary approach to enrich the experience of visual media like in games, and movies: “Auditory cues play a crucial role in everyday life as well as VR, including adding awareness of surroundings, adding emotional impact, cuing visual attention, conveying a variety of complex information without taxing the visual system, and providing unique cues that cannot be perceived through other sensory systems. [...] VR without sound is equivalent to making someone deaf without the benefit of having years of experience in learning to cope without hearing” [56]. Further design studies revealed that soundscapes effect tasting experiences by adding a significant hedonic value [57, 58]. Soundscapes are defined as an “acoustic environment as perceived or experienced and/or understood by a person or people, in context” [59], which means that they represent a sign to their perceivers. We can also observe that, for example, conscious choosing of sounds plays out differently in behavior stimulation of children regarding play experience and the play itself

[60]. Overall, sound design creates imaginative spaces in research and practice and is particularly important for narrative designs [61]. Adopting sound design principles for voice interaction design, we aim to enrich the narrative strength of VAs and explore how this will affect potential users.

3 Conceptualization and empirical investigation of a sound library

Weather reports are a frequently used service of VAs by users. In light of our research questions, we aim to build and evaluate a library of sonicated weather reports as a case study. Thereby, we decided to adopt the approach proposed by Mynatt et al. [62], who discussed potential pitfalls during design and subsequent recognition failures by users during the use of a sound-based interface in their work. In particular, the authors emphasized considering four categories for designing auditory icons: identifiability, conceptual mapping, physical parameters, and user preference. As follows, we discuss relevant theoretical concepts from related fields of sound design. Second, we continue with a user survey to collect conceptual mappings and physical parameters as design materials to empirically ground the design space for sonic overlays.

3.1 Theoretical implications from sound design

Current design practices of VAs focus on advances in speech modulation and interaction while not having established to complement speech-based output with soundscapes, yet. In this context, a sonic overlay can technically be characterized as a second track played in parallel with the voice as the primary track (see Table 1).



Regarding the goal of sonic overlays, two fundamental requirements can be identified that the design should take in mind:

- **Discrimination quality:** As the primary information is given by speech, the sonic overlay must not impede or interfere with the information transmission of the first (talkative) channel.
- **Conceptual mapping:** The second track is not arbitrary but should supplement the first to render the output more expressive and informative.

3.1.1 Increasing the discrimination quality of sonic overlays

In contrast to earcons, the aim of sonic overlays is not to substitute and summarize one specific piece of information but to enhance the experiential quality of information articulated via speech. Therefore, voice and sonic overlays have to be designed in synchronized co-existence to communicate

Table 1 Enhancing the voice track with a sonic overlay

Speech Output	the weather in cologne on monday is sunny
1 st Track Voice	
2 nd Track Sonic Overlay	

and express information auditorily and in parallel. Hence, we take a special focus on what we call the discrimination quality — a category and feature that allows the user to isolate, separate, and process speech- and sound-based information directly.

Krygier [63] has adopted the basic concept of visual signifiers to the auditive channel. He outlines the concept of sonic variables by focusing on abstract sounds that can be modulated by frequency, volume, or timbre to encode information. Studying the variation systematically, he concludes that sound location and volume, pitch, register, timbre, duration, rate of change, order (sequential), and attack/decay are viable sonic variables to enhance geographic visualization. In contrast to Krygier [63], we move the design space beyond abstract sound and consider speech-based output as embedded and discriminable quality of a holistic audio clip. In this sense, Table 2 presents a not conclusive set of sonic variables that aims at the most notable discrimination possible between sonic overlays and speech-based output.

For our design, we took the discrimination variables *Loudness*, *Timbre/Motives*, and *Temporal position* into account which we regard as most impactful in our design. We

discarded the variable *Frequency band* because we aimed for simple and non-modified soundscapes. As smart speakers vary in their technical loudspeaker quality, we neglected to build on *Location* as discriminative quality. However, this dimension might be worth considered in future design studies, as certain listeners using high-end speakers and headphones for VAs on their smartphone, have the technical equipment to experience localization in 3D sound spaces. It might support immersion by, for example, indicating the incoming direction of wind and rain in acoustic weather forecasts. In the following paragraphs, we provide further detail to understand how the chosen variables add to and are reflected in our design.

Loudness Humans can distinguish between different volumes from about 3 dB up to 100 dB. Loudness owns an ordering function by its nature. Keeping a sound experience linear without any variance, loudness might become unconscious over time. Hence, different magnitudes of loudness might highlight and contrast parts of the sonic experience [63]. In particular, different volume levels might increase the discrimination between speech- and sound-based information

Table 2 Sonic variables and their discrimination quality related to voice output

Variable	Description	Discrimination Quality
Spatial Location	The location of the sonic overlay related to the voice output in a two- or three-dimensional space	Effective, depending on the location distance
Loudness	The magnitude of the sonic overlay are related to the voice output	Effective, depending on the volume distance
Pitch/ Frequency band	The pitch of the primary frequency band of the sonic overlay is related to the voice output frequency band	Most effective when the frequency band of the sonic overlay is below or above the voice band (a partial overlapping is possible)
Timbre/ Sound Motives	The general prevailing quality or characteristic of the sonic overlay related to the voice output	Effective when the timbre of the sonic overlay is different from the human voice (e.g., music, abstract sounds, or natural noises)
Temporal position	The temporal location of the sonic overlay is related to the voice output	Most effective when the location is before (intro position) or after the voice (outro position). A partial overlapping and fading is possible)

by lowering the illustrative sounds and turning up the voice volume.

Timbre/motives Krygier [63] defines timbre of sound as the encoding of information by the character of a sound. In analogy, instruments own a characteristic sound, such as the brassy sound of a trumpet, the warm sound of a cello, or the bright sound of a flute. Similar to the human voice, Alexa, Siri, and other VAs have a distinct sound that is distinguishable by the human ear. By choosing and incorporating distinct timbres for sonic overlays, their discrimination quality might be increased. Consequently, using tones or pieces of music, like a bird's flutter or a synthetically produced ambient sound, contribute to recognizing both auditive tracks. This way, information on both tracks can be encoded independently. Additionally, music and sounds transport atmospheres and expressions of emotions, often recognizable as a distinct motive and in movies even underlining principal characters. Such superimposition of motives supports the construction of compound earcons [24] but can also be applied to sonic overlays.

Temporal position By its very nature, audio tracks have a temporal structure and order. Thus, discrimination can also be supported by separating the sonic overlay and the voice track in time. The intro and the outro take a particular temporal position here. For instance, either speech may start or the sound of falling raindrops before the assistant begins talking. Further, incorporated background sounds may support the discrimination of auditive information when speakers pause.

3.1.2 Conceptual mapping: the semiotic of sonic overlays

We aim to create sonic overlays that are not arbitrary but related to speech-based information. The main goal of sonic overlays is to serve as an illustration of what has been said, leading to double encoded information by speech and a sonic overlay. For instance, if the VA reports rain for the next day, the sound of heavy rain supports this information. To characterize the relation between the vocal output and the sonic overlay, we apply Peirce's semiotics [64] similar to David Oswald [65] in his work about the semiotic structure of earcons. The core of Peirce's semiotic is the symbol as a triadic relation between the object, the interpretant, and the sign:

- **Sign:** the sign-carrier which has a perceptual representation
- **Object:** a thing, a concept, an experience, or an emotion the sign refers to.
- **Interpretant:** the perception and interpretation in form of perceived object mood, or emotion in the mind of the perceiver

The sign mediates between the object and the interpretant. For instance, the ringtone of the mobile phone mediates its owner that someone is calling her. In this case, the knowledge of the calling is the interpretant, and the referred call presents the object, while the ring tone is the sign that caused that interpretation. In Peirce's semiotic [64], we can say that the linking of the mobile phone's ringing and its vibration refers to the same object (the call) as well as the interpretant (the knowledge of the call). In the same way, we can now characterize the relationship between the speech and its sonic overlay.

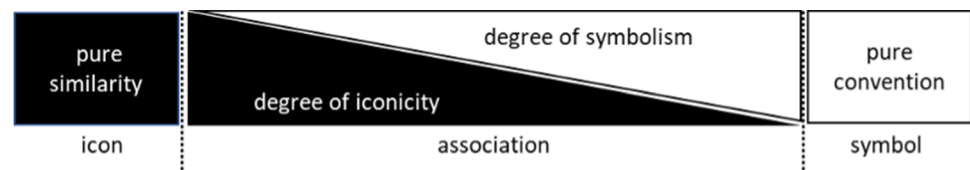
Looking at the encoded meaning in this process of creating sonic overlays [65], Gaver [25], for instance, distinguishes between an iconic, a metaphorical, and a symbolic perceptual mapping. In contrast, Oswald [65] uses the Peircean tradition [64] distinguishing between iconic, indexical, and symbolic signs. Our view is influenced by both authors. Focusing on the experience, we follow Oswald's comment that the constitutive element for iconic signs is similarity, not physical causality. For the same reason, we focus on associations, metaphors, and signal correlations that establish a link between a sign and its object. Consequently, we distinguish between three sign categories referring to the three kinds of relationships:

- **Iconic:** the representation based on the similarity of the signs and the signals produced by the object
- **Associative:** the representation based on associations, metaphors, or correlations between sign and object and the signals produced by the object
- **Symbolic:** the representation based on convention only, no natural link between sign and object

Moreover, we consider this distinction as heuristic classification, where the icon and the symbol represent extreme values (see Fig. 1), when normally a sign has both qualities to some degree: the iconic quality to have semantically and/or signally proximity to the referenced object, as well as the symbolic quality, to draw to the object just by convention and repeated experience. However, we consider that such smooth transitions among the categories will be unproblematic in practice as the primary goal is not to uncover the essence of a sign but sensitize designers about the various opportunities to encode information by a sonic overlay. As follows, we want to discuss the three categories in more detail.

Iconic mapping An icon is a visible, audible, or otherwise perceptible representation of the thing for which it stands. In the auditory world, iconic auditory signs will be sounds that sound similar to the object [65]. Thus, the iconic character results from an imitation of sounds typically produced by the referenced object. For instance, the dog iconically barking refers to the barking dog, or the engine noise serves

Fig. 1 Gradual transition of icon to symbol, from high iconicity to high conventionality (adopted from [65])



as iconic auditory of a moving car. Iconic sound design is typically used in a radio play, movies, and computer games to enrich the user experience. In some cases, weather owns strong iconic representation, like, for example, thunder. We aim further to uncover which iconic sounds and combinations of those are useful to incorporate in sonic overlays.

Associative mapping Going one step further beyond iconic representations, we can uncover associations that are reduced and linked to a distinct characteristic or feature. In the case of Starwars, Ben Burtt looked for familiar animal or machine sounds to establish credibility to ensure recognizable semantics for the sound effects: “The basic thing I do in all of these films [Star Wars and its sequels] is to create something that sounds believable to everyone, because it’s composed of familiar things that you can’t quite recognize immediately.” — Ben Burtt quoted by Whittington [66].

Arbitrariness is based on some similarities between the sound and the referent but not as strong as in auditory icons at the iconic level. As Gaver [25] argues that in general, iconic/nomic mappings are more powerful than symbolic and metaphoric/associative mappings, because iconic/nomic mappings show a direct relationship between the auditory icon and the physical source of the referent.

Symbolic mapping “Auditory icons require there to be an existing relationship between the sound and its meaning, something that may not always exist” [67]. For example, this is the case if weather conditions do not come with literal sounds. A speaking example is the difference between thunderstorms and cloudy weather conditions. Whereas thunder offers an iconic mapping through its distinctive sound of rolling thunder, cloudy weather does not have such an explicit feature. In the absence of an iconic mapping, we ought to apply symbolic mapping, which “is essentially arbitrary, depending on the convention for its meaning” [25]. For example, when the VA announces cloudy weather, the consistent use of a particular sound establishes a symbolic relationship, similar to a ringtone that a user associates — over time — with a particular application.

3.2 User survey design and procedure

The first step in our design of sonic overlays is to define a conceptual mapping that is understandable by the users. Sonic overlays are more recognizable if they are based on iconic

and associative mapping, with an active and purposeful linking between what is said and heard. This has the advantage, that no social conventions have to be previously established. Therefore, we conducted an online survey to collect associations with basic weather report events, such as rain (1), fog (2), frost (3), cloud (4), snow (5), thunder (6), and sun (7). In total, we received 33 complete answers but decided to incorporate also the described associations of 15 incomplete answers. We therefore collected a data set of 48 participants aged between 23 and 66 (male: 12, female: 19, non-binary: 1; mean age: 36.9 years).

Our survey did not aim at being statistically valid since it intended to sensitize our design phase. The survey was distributed in the area of Germany and Great Britain using social media services. All questions were not mandatory and open. We asked two questions for each of the seven weather conditions. First, the participants should name three concepts or terms they spontaneously associate with the mentioned weather conditions. Second, they should name or describe three associations of sound, noises, and/or music. Even if these associations are not explicitly set to music, they give the sound designers an impression of the semantic field that is evoked by each weather condition. Finally, we collected demographic information such as age, gender, and education. Afterwards, we decriptively summarized the results (see Table 3). Therefore, we clustered identical and very similar meanings. The table below shows the 10 most named concepts.








3.3 Results of semantic and sonic associations

3.3.1 Iconic mapping

The survey showed that for the specific weather events, participants had varying difficulty associating tones, sounds, or music, and these associations could be a lot diverse. The association turns out to be most coherent where there is a natural iconic mapping, i.e., where a weather event naturally causes sounds. Rain or flashes represent a fitting example, therefore. In the case of rain, for instance, the associated semantic field revolves around the theme of wetness, water, and raindrops. Those are also associated with certain moods such as chilling, and discomfort but also calmness, and certain colors such as dark and gray.

The theme of rain and raindrops can also be found in associations such as pouring, as well as in associated objects

Table 3 Semantic and sonic associations regarding weather conditions

Icon	Description	Conceptual Associations	Sonic Associations
	Sunny	warmth/heat (23), happy (10), brightness (9), sea/beach (8), ice cream (5), sweating (4), light (3), summer (3), holiday (2), cheerful (2), blue (2), sky (2), yellow (2),	birds chirping (27), songs (14), music types/styles (12), beach sounds/sea sounds (11), sounds of water/splashing (9), children playing (9), laughing/cheering (8), crickets chirping (5), individual instruments (4), high pitched noises (3)
	Cloudy	gray (12), dull (8), dark (6), shadow (4), sad (3), uncomfortable (3), probability of rain or snow (3), sluggish (2), windy (2), overcast (2)	music types/styles (20), sound of wind (9), silence (7), thunder (6), light wind noise (3), light water noise (3), traffic (3), string instruments (2), trees that blow in the wind (2), rumble (2)
	Foggy	mysterious (4), mist (4), damp (4), headlights (3), cold (3), white (2), quiet (2), pea souper (2), epic scenery (2), darkness (2)	silence (12), a dark sounding horn (11), songs (5), muffled sounds (4), slow traffic (3), scary music (3), music types/styles (2), birds chirping (1), crow (1), echo (1)
	Thunder	lightning (9), scary (5), waves/water (4), rain (4), strong wind (4), excitement (4), danger (4), bending/falling trees (3), thunder (3), dramatic (3)	howling/hissing/swishing (9), the sound of thunder (9), whistling (5), rain falling (5), drums beating (5), bangs (5), full orchestra (4), strong, whipping wind (3), sounds/instruments (3), rattling (2)
	Rainy	wetness (20), puddles (6), raindrops (6), water (5), umbrella (4), cold (4), chilling (3), rushing (3), damp (3), pouring (2)	dripping (28), splashing (11), water rushing (6), footsteps in puddles (5), songs (5), pattering (3), drumming (3), music types/styles (3), opening umbrella (2), cars going past on wet roads (2)
	Freezing	ice cracking (18), coldness (13), white frost (6), freezing sounds (6), ice (5), slippery (5), danger (5), dressing up warm (4), clanking (3), single instruments (4), snow (2),	music types/styles (8), crunching (5), ice skates on ice (5), shivering (4), scratching (4), scraping on cars (4), slipping and falling (2), songs (2),
	Snowing	white (11), cold (10), snowman (7), brightness (5), silence (5), calm (4), winter (3), snowballs (3), winter sports (2), flakes (2)	(snow) crunching (16), Christmas/winter songs (11), silence (12), ice skates sliding on ice (3), Christmas music (2), clumping (2), shoveling snow (2), crackling fireplace (2), soft music (2), muffled noises (2), bells (2)

The numbers in brackets refer to the number of participants mentioning the respective association

for personal protection, such as an umbrella. The sonic field translates the theme of the semantic field of water in terms of nature sounds caused by rain, e.g., splashing, water rushing, dripping, pattering, and drumming. Besides, mainly naturalistic or nature-simulating associations were named, e.g., thunder and lightning, running faucet, or rice grains weighing back and forth. In the case of lightning, an iconic mapping is found in most cases that the electrical discharge produces not only lightning but also thunder. This fact has led to a quite homogenous sonic field, where most participants directly associate thunder or specific forms of thunder such as dumpling, crashing, or banging. Furthermore, it turns out that lightning is semantically and sonically associated with rain, expressed, for example, by sonic associations such as drumming, waterfall sound, or pattering rain.

3.3.2 Symbolic mapping

The opposite case presents weather events where a natural iconic mapping does not exist. The most prominent example of such a case is the cloudy weather. In contrast to rain or flashes, the participants do not associate specific natural events or activities but a vague, general impression of gray, dark shadows, coldness, and a quite unspecific, melancholic mood of discomfort, sluggishness, and bad temper. The theme of coldness was also expressed by mentioned protection means like bringing a jacket or sweater weather. Occasionally there are associations with seasons, e.g., autumn or places like Germany, as well as activities such as doing sports outside or city trips. This broad, unspecific, and, as it were, the soundless semantic field is echoing in

the sonic field. Here we observe the heterogeneous answers that aim to differently translate the vague ideas of gray, gloomy, and melancholy sonorously. In addition, two participants answered the question about associated concepts but omitted the question about sonic equivalents. The other answers show a wide range of sonic associations. What is striking here, is the frequent tonal characterization of cloudy by general musical characteristics (melancholic music, ponderous beat, polyphonic male choir), certain musical trends (lo-fi beats, jazz music), or individual instruments such as strings, and styles of sounds such as muffled sounds or dull hum without associating one specific sound or piece of music. Participants mentioned natural sounds such as wind or water less frequently. We also find it inspiring that some participants associated human noises like sighing, breathing, and the sound of yawning to give the melancholic mood a sound.

3.3.3 Inbetween iconic and symbolic

The answers further show that most associations cannot be unambiguously classified as iconic or symbolic mapping, but mostly represent something in between. Therefore, in our view, it makes sense to understand the schema outlined in Section 3.1.2 as a heuristic rather than a strict category system. Sunny weather is one of the examples where iconic, associative, and symbolic mapping is balanced. In the semantic field, we see strongly iconic responses, e.g., warmth/heat, brightness, and blue sky, but various answers more indirectly related to sunny weather such as summer, expressions of summer feeling like being motivated, happiness — for instance, expressed by laughing — as well as diverse summer activities such as cycling or eating ice cream. In addition, some answers refer to measures for sun protection, e.g., sunshades or sunscreen. The corresponding sonic field also reflects this semantic field. Unlike flashes or rain, the sun does not directly cause sounds. The associated natural sounds are not iconic but rather associative. Various participants mention sonic expressions typical for a sunny sea holiday, such as the sound of waves, splashing in the water, or voices at the beach. These associations present indexical signs in the sense of Peirce [64] because of the causal chain of the sun (causes hot causes refreshing beach holiday causes sea sounds).

By the same token, they present a metaphorical mapping in the sense of Gaver [25] because in western societies beach sounds become a metaphor for a hot summer, good feeling, and sunny weather. In addition, some participants associate sunny weather with crickets chirping or birds chirping. Again, there is an element of indexicality and metaphoreness in these associations (as not sunny, rainy weather physically impedes both, chirping and singing, and so both natural events have become metaphors for a sunny summer). Less indexically but more metaphorically are answers such as

laughing or cheering. While both are not directly metaphors for sunny weather, they are metaphors for happiness and good feeling — which was one of the associations in the semantic field. This feeling of lightness, sunny weather lifestyle, and good mood are represented by many musical associations, both regarding styles (light pop music, light electronic music, major sounds, reggae, Latin American music, as well as regarding particular songs such as “Sunshine Reggae” from Laid back, “O.P.P.” from Naughty by nature, and two ice cream commercials, “So schmeckt der Sommer” (Engl. “This is how summer tastes”) and “Like ice in the sunshine”).

Overall, the answers indicate that iconic mapping is prominent when the weather event causes typical, easy-to-remember sounds. In contrast, when those sounds were not available, participants suggested symbolic mapping more often.

4 Developing a library for sonic overlays

Our library for sonic overlays is based on the empirical and descriptive results of the survey described in Section 3.2. Further, we use the categories of iconic, associative, and abstract sound to cluster the results and produce sound clips that show a high discrimination quality for all seven weather types. We will explain our design rationale and according steps as follows.

4.1 Design approach to enrich Alexa’s weather report

In contrast to Mynatt et al. [62], we decided to gather conceptual mapping and physical parameters by a free-form survey before the design phase. Further, our goal is not to design auditory icons but to illustrate speech by using iconic, associative, and abstract soundscapes that are not synthesized into an identifiable sound-only design but serve the purpose to illustrate spoken information.

The seven most distinctable weather types were chosen to be the core of this design: sunny, rainy, cloudy, foggy, snow, frost, and thunderstorms. The authors sorted the responses into categories depending on each sound’s connection with the weather in question:

- Iconic sounds, which are caused directly by the weather
- Associated sounds, which are expected to occur in conjunction with the weather but are not directly caused by it
- Abstract sounds, which have a connection to the stated weather type in the respondent’s mind but are not necessarily linked to it

Table 4 Sorting and categorization of survey results using the example for rain

Iconic	Association	Symbolic
Rain noise	Footsteps in puddles	Car horns
Dripping	Tyres on wet road	Many voices in closed space
Splashing	Opening umbrella	Drumming
Rain on the roof	Rustling raincoats	Boiling water
Storm noise	Wind blowing	Tapping
NA	Drains gurgling	White noise

This categorization is based on previous conceptual considerations, as introduced and explained in Section 3.2, and enables easier identification of any positive or negative reactions to certain types of sound by users. Further, we want to highlight that, in contrast to rain, certain weather conditions like foggy and cloudy have no iconic sounds. This must be taken into account when creating the respective soundscapes. It will also provide an opportunity to evaluate how a lack of iconic sounds affects the user's overall perception of the soundscape.

Therefore, as a first step, we categorized the survey results described in Section 3.2, sorted from most common to least common, for all seven weather types (see Table 3). Table 4 exemplifies a rainy weather condition (see below).

4.2 Structure and elements of sonic overlays

Sonic overlays and earcons/auditory icons share multiple features, such as conceptual mapping and encoding information by sounds. Yet, the survey of Cabral and Remijn [68] shows that in contrast to sonic overlays, earcons are quite short (mostly between 0.5 and 3 s). As our sonic overlays attempt to illustrate speech-based information of VAs, we need to take into account that talking often lasts from a few seconds to minutes. For instance, the weather report of the German Google Assist takes about 10 s, allowing sound designers further options regarding rhythm, using pauses, proving ambient sonic overlays, and other temporal parameters. Another main difference is that in sound overlays, the voice conveys the primary information, which liberates sound designers to more subtly encode the information and, for example, emphasize or ironically comment on the spoken information by sound. However, it also creates new constraints, such as that the sound overlay should not interfere with the voice making it difficult for the user to understand what the assistant has said.

The examples created were each around 25 s in length and incorporated sounds based on the most frequent answers given in the survey, in combination with a synthesized voice similar to that which would be heard from a VA. Further, a proper difference in loudness between the soundscape and speech ensures the discrimination quality within the sonic

overlays. The structure of each sound overlay clip was consistent across all the weather types: each starts with around 5 s of sound effects to build up a soundscape representing the weather, then a voice would explain the weather condition and temperature, followed by additional 10–15 s of audio. If the clip includes any musical elements, these are incorporated into the soundscape after the voice has spoken.

Musical elements and soundscapes are essential to creating an expressiveness of information that speech could not. Two examples also incorporated musical elements, besides sounds and spoken words. The example for the sunny weather condition incorporated a guitar melody inspired by “Here Comes The Sun” by The Beatles, as this song was mentioned by multiple survey respondents in association with sunny weather. Further, the example of the frosty weather condition incorporated an original melody using tones and timbres identified in the surveys as conveying a feeling of cold, icy weather. When creating the soundscapes, sounds with a rather direct connection to the weather type in question were prioritized, e.g., the sound of wind or falling rain. However, in impossible cases, more abstract sounds were preferred instead, e.g., the cloudy soundscape that featured heavy traffic noise. In either case, all sounds featured in the soundscapes were selected from the survey responses.

5 Evaluation of the sonic library

5.1 Interview study design and procedure

Frequently, associations and imagination are linked to prior experiences and their cultural background [23]. Therefore, we did not aim at a statistical representation of the populations in Germany and Great Britain. We were looking for participants with heterogenous cultural backgrounds able to speak and understand the English language. For recruitment, we used snowball sampling in our extended networks [69], thus, we posted requests in social networks like Facebook, international telegram groups and private messenger services. To further diversify our sample, we asked the first participants for references from their extended networks. Most of the 15 participants (4 male, 11 female), currently lived either in Germany or the United Kingdom, in addition to one participant living in France and one in Palestine. However, their geographical backgrounds were significantly more diverse, including south-east Asia, Sri Lanka, Canada, and Russia, among other countries. This diversity in backgrounds helps identify how a person's current or past environment might affect the evaluation of sonic experiences and weather types. Table 5 provides an overview of the corresponding data regarding age, gender, and current and previous residence. Most interview participants had at least some previous experience with VAs. Participants who were inexperienced

Table 5 Study participants (n=15) representing international differences in culture and residence

ID	Age	Current residence	Additional info	Previous residence
P1	27	Germany	Industrial, edge of forest	Hong Kong
P2	29	UK	South England, rural	/
P3	23	UK	Hull, suburb	Used to live in a more rural area
P4	62	UK	Scotland, coastal	Used to live in New Forest
P5	57	UK	South England, countryside	/
P6	60	UK	South England, countryside	/
P7	28	Germany	Industrial, edge of forest	Azerbaijan
P8	25	Germany	Industrial, edge of forest	Toronto, Canada — less rain, colder in winter
P9	67	UK	Guildford, leafy suburb	Sri Lanka
P10	20	France	Small town, near coast	/
P11	23	Germany	Rural, edge of forest and small town	Spent 3 months in Canada
P12	25	Palestine	Varied seasons, hot in summer, rainy winter	SE Asia — weather very similar all year
P13	46	Germany	Small mountain town	Village in Lower Saxony
P14	33	Germany	industrial area, city	St. Petersburg, Russia
P15	31	Germany	Industrial, edge of forest	/

in interacting with VAs had a basic understanding of how they work. Therefore, we only explained the sound overlay concept. Participation in our study was voluntary and did not involve any compensation.

We chose a qualitative interview study approach to explore the subjective perception and usefulness of the sound overlay library. Each participant listened to both conditions: VAs with speech only and VAs featuring speech with sound overlay for three randomly chosen weather types. We created a randomized experimental design without repetition, so that each participant was played two of the three sounds, e.g., weather report with/without sound overlay for rain (1), fog (2), frost (3), cloud (4), snow (5), thunder (6) and sun (7). First, randomization without repetition ensured that at least six subjects listened to each of the seven weather reports. Second, the randomization was intended to minimize a sequence or order effect. The experimental design randomized the order and also the combination of the other samples (e.g., with snow and storm or with frost and sun) to account for possible changes in opinion brought about by hearing particular examples in combination. Additionally, the order of the clips for each weather was also randomly selected, taking into account that listening to the first clip might influence the next. We uploaded the sound library to youtube to share only the chosen links to the clips during the interview. After listening to each clip, the interviewee was asked specific questions about what they had just listened to, followed by more general questions about the concept and their impressions of it, e.g., did you recognize the sound as the correct type of weather? How long did it take? Or did the information come across,

and how does it make you feel? Each interview lasted around 35 min on average and was conducted over Zoom.

Finally, the interviews were transcribed verbatim and coded inductively and independently in MaxQDA by two researchers using thematic analysis [70]. We focused on the effective sonic experience of the weather types and the perceived differences in design and usefulness. Also, we explored the impact of combining speech and sound and its implications for structuring and contextualizing information.

5.2 Findings

Some participants regularly used VAs to check weather forecasts but the majority relied on websites or smartphone apps instead, usually citing the level of detail offered as the reason why. Several stated that the short spoken summaries by VAs did not give enough specific detail to plan a whole day.

5.2.1 Supporting imagination and experience

Sonification aimed to support people to produce images in their minds that use emotions and prior experiences associated with distinct and ambient noises. By using the examples of weather, we could observe clear challenges in design for two specific groups of weather types: almost silent events like fog, sun, frost, and cloud, and loud events like rain, snow, and thunder. Although the prestudy foreshadowed possible challenges to design recognizable and unambiguous soundscapes, the cloudy weather seemed to cause the

majority of problems in correctly understanding the presented information.

Most of the participants responded to the idea positively when listening to the samples and expressed vivid accounts of their imagination. Some welcomed their emotional responses and explained that this makes the interaction less boring and monotonous but more dynamic (P1). This evokes a space “like being on a boat in the ocean” (P3), when listening to the audio clip of “fog”. According to P1, weather reports supported by soundscapes felt less “artificial” than speech-only and created a kind of “haptic feedback” of the information:

“I think it’s more emotional because you do have like, an image, sort of, in your mind. Yeah, I like the fact that it’s not only rain. It feels like car and rain or some background noise. You know, it feels like you are really in the middle of the city. And you don’t have an umbrella, and you are suffering from a pool. (...) In this context, I think you want to use a temporary, really precise message of the weather, and I think this achieved their goals.” — P1

In particular, the soundscapes emphasized typical feelings associated with specific weather conditions, as participants explained that the thunder sounds made them anxious (P3), a sunny city equaled good feelings (P2, P7), or freezing temperatures indicated not to go outside:

“So we were like heavy winds, which were full of crystallized snow. And you could hear yourself like walking through the huts. Cold, like the freezing or the snow, which feels like the ground. And, yeah, the wind was so strong that you did not want to go outside at all.” — P14

The soundscapes of pleasant and unpleasant imagined situations alike enhanced the intended message and supported possible adaptations of the participants’ behavior, like being motivated to go out (P8). Some saw the concept particularly useful for special occasions and ambient background information needs (P13). Moreover, P1 and P14 reported that the sonic overlays contributed to a calm and relaxed feeling.

“Natural sounds in general. Also the crows and animals and things like that. Because sometimes people are stressed about everyday life or life pretty often. So they have, they want, like something to relax. And maybe one selling point of this app or a voice assistant would be like that one can relax, that are in our everyday life.” — P14

Sound is not considered overall necessary for a system solely designed to give factual information (P12). While regular forecasts are unbiased, sound adds a character to it that can have positive or negative connotations. This can help

to form decisions based on the weather because it is easier to imagine yourself in the context. P12 indicated that the specific information might not be as memorable, but the overall impression was much stronger and helped with understanding the consequences of the weather conditions. Another piece of feedback from several participants was that the soundscapes made it easier for them to visualize the weather and think of how to prepare for or react to it. P3, P11, and P10 considered this useful for morning routines or directly after waking up in a dark room. Moreover, P10 calls the design concept more reassuring by giving a feeling of naturalness and coziness (P10). P3 also was surprised that it was not already commonplace for VAs since visual apps use graphics to add more context and to communicate information in a more appealing fashion (P6).

Further, this concept bears a chance to give friends coming to visit a more precise idea of the weather conditions and makes it more interesting to share (P13). Additionally, it might help to feel a deeper connection and experience with the represented location if you live far away, as long as the information represents the reality:

“Let’s say I want to go to London and I’m checking the weather in London. Or maybe I want to see the weather in a different country right now. For a particular reason, it is important to me. (...) but instead of saying rain and the strength of the rain, it might add more because if it is on real-time as opposed to a forecast, if it is music, then I feel it. This level of, you know, the burden of interpretation. But if they are actual, it’s almost as if they are giving real-time Information. Then if they are making me hear it, how it is, how snow is flowing. They know how it is raining in London or wherever away from the I can see from my window. I can see data that has been an interesting dimension that I would be interested to see.” — P9

Meanwhile, missing experiences of weather conditions or landscapes might contribute to misleading interpretations or less precise perceived information. For example, P15 could not recognize and relate well to the foghorn sound that represented foggy weather in comparison to P4, who imagined their current residence:

“I could picture the coast where I live, which is a harbor, small harbor and the sea and foggy sea and the fog coming into onto the land, which it does where I live (...) quite often. So yeah, a totally foggy, virtually visible. With the emphasis on sounds that you hear rather than what you see.” — P4

As P1 grew up in a large city, hearing footsteps in the snow made it difficult to differentiate between snowy and frosty weather and carried over all the impression of a hiking

vacation in nature rather than an intuitive sense of the weather conditions. She was missing the noises of traffic, for example, cars. In contrast, P6 noted not to include traffic noises because those do not symbolize sunny weather to her. In a similar vein, P8 and P13 did not consider children playing outside as an appropriate illustration of sunny weather, and hearing splashes reminded P13 rather of rain.

5.2.2 Sonic information design

The sonification of information relies on abstract and iconic sounds, as well as relevant music pieces and speech. Particularly abstract sounds contributed to an active imagination and conveyed the meaning of the weather conditions. Therefore, all participants pointed out that the incorporation of related sounds gave a better impression of the scenario:

“I think all of these have given me very if you’ll pardon my illusion, Animal Crossing kind of vibes. I don’t know if that was a deliberate image or just circumstantial. But it’s not the weather. The tones fit the weather, the sounds of the light. With this one, you could hear like it was like birds singing. Nice day, kids having fun. Like, I think that was a roller coaster. And then the marimba at the end or like a guitar.” — P12

Overall, the concept does not represent a simple sequence of symbolic sounds. Hence, the soundscape has to be layered with consideration. An urban environment might sound different than pure nature but it has an equivalent impact when sounds like background noises are combined that indicate events happening during this kind of weather or the place of experience.

“I like that. Not just the sound of it. It really sounds like you try to mix it with different elements like the surroundings. Sometimes the sound is not really directly about the weather, distinctive. But I think that’s really awesome. Some feedback is that, for example, there’s the second one I have the most problem understanding. The foggy one.” — P1

The participants appreciated incorporating musical elements that acted in a similar vein to convey information and emotions that noises could not. For example, P11 stated that music represented “icy” conditions much better than footsteps. Likewise, this type of sonification supports the differentiation of similar states like frost, ice, and snow. P2 explained that music was thematic and indicated light and pleasant snow by that:

“I think it was very thematic in the sense that it gave you an idea of what to expect. It kind of indicated it’s going to be like, you know, sort of like, oh, it’s nice. You can walk in it. It’s going to be like pleasant thunder.

It didn’t seem to be indicating snowstorm: Stay in your house!” — P12

Likewise, the use of a guitar, for example, may produce a “calming effect” (P8). In contrast, P11 described VAs as a convenience and aimed for efficient interaction, where music might be in the way. Further, P12 was concerned that not everybody would appreciate such a design decision as well:

“I liked it. I mean, again, it’s I think the sort of people that would be put off by the extra fluff at the end. People that would just look at a website and wouldn’t use the service anyway. So I think it’s adding an additional level of sort of engagement to people that are going to be using the product.” — P12

However, the music proved to be an effective element for supporting imagination and speech-based information:

“All the right information came across straight away. And what was interesting was that because I’d heard the music first, I had this same image of this road going into the distance and everything, a little bit orange. Don’t ask me why, but maybe going into the sunrise, sunset, you know, a pleasant travel image, basically.” — P4

An overall trend in the results was that soundscapes that more heavily used iconic sounds were more well-received than those which relied solely on abstract sounds. This presents an issue for weather types that do not have any associated iconic sounds, such as cloudy or foggy. Especially iconic sounds are well suited to represent precise information, entail clear messages, and evoke past experiences as associations at the same time. Further, natural sounds are closely tied to the expectations of weather conditions:

“And because of the sound of the birds, you kind of feel it’s sunny and the kind of feel that people outside and that things are happening outside. So you assumed your kind of mental image was this sort of like sunnier, drier weather.” — P9

In comparison, particularly rain and thunder were tangible noises with high and quick recognizability. Participants (P13, P3, P2) discussed afterward, for example in the case of snow and frost, how the granularity of weather conditions and their differentiation could be supported by a variety of iconic noises.

“And as I mentioned before, you could play a different thing. So the severity of it. So you’ve now winden and instead of sort of a lighter sound, but more heavy, I assume they were sort of sleigh bells or reindeer to indicate a more hazardous conditions maybe. Yeah, but yeah, I know it was all very easy to hear that it gave across everything you were trying to say.” — P2

P13 added that it could be confusing if there are snow sounds but only 50% chance of snow, for example, and that it may be better to build up from a wider bank of sounds for variations (P3), for example, a concise representation of temperature and that “Rain sounded maybe not as ‘heavy’ as the voice said” (P3).

Besides, difficulties arise with sounds that cannot be represented iconically because of the absence of noises, for example, with sun, clouds, or fog. However, this might lead to confusion by trying to substitute by using crows or horns that occur or are used in cases like fog. P11, P10, P4 and P1 had trouble understanding the meaning of crow noises and considered them as confusing.

“Then I think they were crows or rocks, the birds. For me, they could make that noise. Morning. Evening. Any weather? Probably. But then not everyone’s going to know that I live on the coast. And for me, I was wanting to. Seagulls, of course. But of course, it’s not everybody lives on the coast. So, yeah, it wasn’t a big deal, but the phone comes with a real positive clue, so it didn’t matter. The rest was just atmospheric. Quite nice to listen to.” — P11

At times, some participants (P8, P1, P9) felt overwhelmed by the combination of too many sounds and suggested cutting back (P8). Musical elements could of course be added but also detracted from the message and would leave just a noisy impression (P9). Overall, the balance of iconic and abstract sounds provided an enhanced experience and emphasized the information. Nevertheless, the design should focus on communicating a clear message as well:

“I liked that they didn’t all do the same thing. So you had some that were the literal sound of the weather and some that with sounds associated with the weather. I liked that there was a bit of a mix. I didn’t like that, I didn’t feel like any of them gave a clear communication of temperature. (...) I liked the sounds there and I liked the length of them.” — P3

5.2.3 Adding sound to speech

Besides iconic sounds, a deliberate choice for the design of sound overlays was to incorporate speech providing precise weather information. Many participants claimed that without speech, they could not identify the correct weather conditions, especially concerning fog, frost, and clouds:

“Well, what I noticed is that the abstract sound only came after she talked. The voice (...), there was no ambiguity. And I really knew that it was the frost that made the sound.” — P11

In contrast, some participants indicated that in the case of rain, the speech felt even unnecessary, and, in the case of thunder, it was even more clear than vocal information:

“I felt like it basically brought things across. The voice said heavy thunderstorms. And I feel like maybe the rain wasn’t heavy, heavy, heavy. But at the same time, that would raise the question of, well, how many different words does a voice assistant use when describing weather? And then can you map all of those words on to a sound of rain, like the thunder sounded heavy?” — P3

Overall, the intended and sonificated meaning of rain, sun, snow, and thunder was recognized most frequently and almost immediately. P11 added that by the sound he imagined, it is even easier to remember to bring an umbrella. Further, P12 explained by listening to thunder that he had clear thoughts on the preparation for the upcoming stormy weather.

“I think it was like supporting the voice. Sometimes I also think that the voice was completely unnecessary. In extreme beavers and extreme weather conditions, for example, when it was like snowing or raining. But a service (...) it will be like necessary to at least say the temperature. And I mean the information about that it’s snowing.” — P14

However, most participants considered speech for quick and precise information, like temperature indications (P14), valuable, especially those participants who might be impatient because they are in a hurry (P14, P1, P9). Furthermore, participants feared that voice and soundscapes could compete for attention sometimes, e.g., because of false expectations regarding the timed structure:

“Since, I think, it’s one minute. Whenever, (...) it’s not necessary, but it can be of it can be a bit frustrating if you missed the moment that it starts saying.” — P10

Further, P10, P13, and P12 expressed concerns that voice and background noises were overlapping too much, e.g., children screaming while playing outside (P13). Hence, despite a better image of a complete scenery, speech-based information was drowning down:

“In the same instance, you get like in films, sometimes there’s a dialog scene. And then the orchestral score or the things in the background is so loud, you actually can’t hear what’s going on, which then detracts from the product, which I think is something you guys have managed to avoid.” — P12

Additionally, P11 mentioned that sound shouldn’t seem to contradict speech to not add to ambiguity and confusion:

“It doesn’t add more information to this, to the stuff that she’s saying. Because in the first part of the snow, it added snow. She didn’t say anything about snow. And the second one added wind, even though the voice just said it’s foggy, not windy. And it must be very difficult to achieve. But I think that’s really important that the sound is very much in line with the words and not adding or taking away information.” — P11

P11, P8, and P9 stated that the use of sound elongates the application and requires patience. Consequently, in their need for quick information, they would prefer speech-based, either through voice or by glancing at their phones.

“In a car. Probably like when you need to just have the information (...). But when my mind is like, I just want to know this and then I want to do something else. I don’t know in which situation that’s the case. Usually, most of the time, but when I ask: ‘Okay, what’s the weather going to be like?’ And then they tell me and then I cannot ask another question for like 5 seconds because I have to wait until the rain stops. That would annoy me so much.” — P11

In total, we could observe balanced opinions on the preference of voice or sound—first regarding the structure of the sonic overlays. Therefore, some of the participants (P6, P14, P9, P11, P5) argued, for example, P14 and suggested starting with speech first when designing sonic overlays:

“I think it will be better to start with a voice or maybe a millisecond off or a nanosecond. I’m not sure of like of forever, of a silence and then the voice. Because I think sometimes people don’t have patience. Some people don’t have the patience for waiting until the voice pops up.” — P14

P6 demanded to have speech instantly — “facts not thrills” — but could imagine maybe a short sonic fade-in before and fade-out quickly afterward. A further advantage of speech—first might be reduced ambiguity and sound as additional layers that can be better interpreted (P9). P11 suggested making the clip shorter overall to make it more efficient, although this might lead to impressions that interfere with the voice.

“Waiting in suspense for the voice - then it happens suddenly. Voice and sound should start at the same time then let the sound carry on for just a few seconds afterward to leave an extra impression.” — P11

On the other hand, participants had found reasons to start with sound as well:

“No, I think the fact, that the lead-in was an audio clip of the weather type or something alluding to the weather type followed by the information, then followed by

another weather clip with a bit more music. I think it gave you an idea of what was coming. It was then clarified and then you got this sort of little ribbon on the top of whatever you’re referring to us.” — P2

Many participants appreciated the current design structure of the sound overlays. They pointed out that sound introduces impressions and scenes as afterward speech fades in to confirm and clarify weather conditions. Besides, P10 describes this design as feeling less aggressive than the assistant speaking at you immediately. Nonetheless, participants like P14 and P4 emphasized that this concept needs time to get used to it first.

5.2.4 Sonic contextualization of experiences

The sonification of information might be extended to other applications and design spaces, as the statements of our participants show in the following. However, they expected some limits regarding the usefulness and experiential value. In particular, situations that allow for ambient sound and personal moods that welcome entertainment, e.g., driving in the car or waiting in general. For instance, P8 considers background ambiance, like the sound of a fireplace or ASMR (Autonomous Sensory Meridian Response) for cooking or studying as relevant. P13 would consider hearing the sounds of frying/chopping, etc., to be more amusing. Additionally, P4 describes a possible situation at work:

“When I’m working in home office, I’m able to choose. When I go out for a walk, I could look out the window. But in Scotland, that won’t tell you. You really need to know that temperature, preferably what it feels like. I mean, that’s peculiar to Scotland. It doesn’t set up. The temperature is what you really want. And yes, I could come out to whatever I’m writing or reading. And I could click or met Office, and I could get it. But if I could just get it instantly, you know, like that just: ‘Oh, I wonder if I need a hat and a scarf as well as a coat today. Do I need two pairs of gloves or one?’ Then I would quite like that. A fun way of doing it, especially as I want to then forget about work, although I actually associate my laptop with work. So for me, just to have some quick little sound, and off I go for my walk” — P4

Besides asking for the weather or specific information, the news is a frequently used service of VAs and radios alike. However, our participants had contradictive thoughts on the sonification of this offer. P4 could imagine a benefit of applying sounds to the presentation of traffic updates, travel reports, or election/sports results, especially at times you want to know the info in a flash. In contrast, P1 expressed

that sound might distract or manipulate information. Further, for P3 bad or scary effects might be reinforced.

“Honestly, I, I don’t, I cannot think of anything that would benefit from that. Because it always conveys some sort of interpretation or maybe opinion or emotion. So if you add it to a news article, it’s not neutral anymore. And I read the news to make my own opinion. So I wouldn’t like to be presented with somebody else’s emotions.” — P11

Whereas more participants can see potential design spaces at home by enhancing other media and smart home applications. P3, for example, would wish for audible feedback on loading times and completion of tasks. P1 explains in further detail how a sound or earcon library of a current VA might enhance the notification experience of deliveries:

“Alexa might have some sound ding ding on this topic. Another possibility is when I’m anticipating a package, I know the different stages of the package, like, is it a ship that is delivering (...). It will be quite helpful because right now, they treat it as a notification. Like maybe you have, you can extend these to some parts of: ‘Are going to arrive today’. If they can have a different sound to describe where exactly my package is.” — P1

6 Discussion and implications

In light of our research questions, we want to discuss our results and provide implications for the design of future voice interaction. So far, Alexa is seen as Voice Assistant, very neutral in their answers with little capabilities to express emotions [18]. A significant amount of research in the fields of speech science aims to address this shortcoming, respectively emotional speech and voice design [7–10, 14, 39, 40, 44, 45]. In this paper, we complement this area of research by outlining a supplementary approach, using sound as a modality that could add a new dimension to voice interaction and enrich the user experience. In particular, we focus on the relation between speech and sound and the balance between communicating information and inducing emotions through sonification.

6.1 Sonic encoding for voice interaction design

6.1.1 Building soundscapes

The prevalent design paradigm regarding sound is to precisely encode information to substitute functions and representations [24], leading to different kinds of auditory icons and earcons that are highly recognizable. However, that also requires either a clear sonic representation, or users to learn

its meaning first. As with current VAs and computer systems in general, we can observe the use and purpose of earcons to signal warnings or direct attention to events on short notice [24, 25]. However, iconic sonification might come at the expense of rich soundscapes capable to transport emotions, atmospheres, and further experiential qualities, as known from the design of classical media and extended realities [22, 23, 56, 57].

Extending the purpose of sound by substituting single functions and representations, our results indicate that sonic overlays may support voice interaction to encode, illustrate and communicate messages. The combination of iconic, abstract, and symbolic sounds shows a positive impact on the perception of weather reports by speech-based interaction. Participants described their experience as stimulating and entertaining, quite the opposite of previous experiences with VAs. Thereby, iconic elements support the recognizability of intended messages. Some weather types gained noticeably less positive feedback than others, particularly weather types that relied more heavily on abstract sounds such as cloud and fog. As these require the listener to draw connections between the sounds and the weather in a less direct way, they are more open to interpretation and have more potential to cause confusion. These potential issues first appeared as early as the pre-survey; these weather types had fewer associated sounds suggested overall, and the most common response for a sound associated with fog was “silence”. Musical elements as well as abstract soundscapes serve as an illustrative layer to build a holistic impression of the specific weather conditions and are a carrier for moods and emotions. However, a missing combination of iconic sounds might obscure some information.

With our work, we present a structured design approach to sonificate and illustrate voice interaction and, thus, enrich the experience of weather reports. So far, only a little work on methods and research regarding design approaches of voice interaction, especially in combination with sound design, exist [20, 21]. Current approaches to voice interaction design are based on collecting example dialogues, spoken terms, expressions, and paths as design materials. Similarly, we collected associative mappings for each message of a weather event and categorized those into abstract, iconic, and symbol design elements to develop a not exclusive sound library. Although the design was well appreciated, we need to balance abstract soundscapes that affect the experience with iconic sounds, meanwhile ensuring recognizability of the intended message to communicate information successfully.

6.1.2 Layering sound and speech

As our results indicate, the sonification of interaction opens the design space for more ways of expression [20–22]. However, voice remains a precise channel to communicate

information and is perceived as an efficient and convenient way of interaction. Therefore, participants expect sounds to illustrate exactly the information of the voice channel and avoid contradictions from both channels. Further, by using abstract concepts like “children playing outside in the sun”, designers have to be careful not to mix channels in parallel that entail soundscapes based on human voices. Otherwise, the discrimination quality is not guaranteed. Besides, more research into differences in similar weather types like frost and snow could prevent misunderstandings. However, participants were skeptical whether, e.g., 50% probability of rain, could be communicated via sound. Yet, they still desired a high granularity to express the characteristics of weather conditions.

The structure of the audio clips regarding the temporal position of sound and voice received mixed feedback from the participants. Some liked the structure of starting with the sound, then introducing the voice, and ending with more sounds as it gave them time to form an impression of the weather from the sound that later was confirmed and clarified by the voice. However, other participants felt that the clips in their current form were too long and that they wasted too much time compared to a voice simply speaking the weather forecast in just a few seconds. Although almost all believed that the sounds produced a better connection to the weather than the voice alone, several interviewees indicated wanting to hear the voice-first to get the most information as quickly as possible. However, a more matching combination of both might reinforce the impression that the sound illustrates what the voice was saying in real-time. Currently, the voice simply speaks over the soundscape after a few seconds.

Overall, sonic overlays illustrated and strengthened the voice message. Speech added the preciseness of information, especially for events or impressions that naturally are silent and hard to sonificate. Besides, a certain granularity and discrimination quality in sound design might positively impact the preciseness of information. However, the temporal position of sound and speech has to be purposefully integrated into the overall design and needs more research to give clear implications.

6.2 Balancing emotion and information

6.2.1 Authentic soundscapes

Data sonification may serve both purposes, conveying information and emotion [22]. Sound design in Science Fiction gives the future a voice, linking the effects to the imagery to enhance the credibility of the cinematic reality [66]. The same holds for the role of sound design in games and XR [56]. Oftentimes, the goal is to create new worlds and experiences that are not nonexistent or less prevalent in real life.

This was quite the opposite for our study because participants expected to understand the sonic overlays effortlessly. The main goal shifted to imitate the surroundings of known places and build on past experiences to encode information. As our results indicate, social context and personal residence environment greatly impact the upcoming associations and respective interpretations. For example, people who live in big cities might practice hiking as a seldom leisure activity, whereas people from the countryside might have a distinguishable understanding. The same applies to cultural experiences, e.g., festivities like Christmas associated with specific music and instruments. However, besides supporting the imagination of the known, places in different parts of the world can be illustrated in the same way. Yet also, in this case, it might be perceived as more worthwhile to experience representations quite close to the original experiences of people living in those areas.

Finally, experiences could be even further personalized by using location data, information on the surroundings in this area or during the daytime, and other chronic data to match the experience of the area. This approach would allow for enhanced recognition of sonificated information and for users to empathize with new places and experiences.

6.2.2 Encoding emotion

So far, VAs lack an engaging experience that motivates users to interact on a regular basis [18] and are regarded mostly in utilitarian ways by users. Following the call of researchers to explore potential experiential qualities of VAs [20, 21, 71], speech science research [7–10, 14, 39, 40, 44, 45] aims at encoding emotional information and expressiveness into the sound of voice and the way of speaking. With our alternative design approach, we investigated the design space to develop and promote an expressive context for dubbing, voice-overs, and future voice acting [72, 73].

Furthermore, our study focuses on exploring the various options to design surrounding and ambient sound contributing to the affective experience of VAs. Our results indicate that sound overlays could enhance imagination in comparison to voice-only interface design. Moreover, our participants reported both calming and anxious effects that either feel relaxing, or symbolize and promote action. This is also due to sound building up a closer complete scene, making it easier to visualize and respond than simply hearing words.

In the tension field of expressive and informative interaction, designers act responsibly and consciously regarding the sonification of positive and negative experiences. As our data shows, some participants were concerned about manipulative misuse of sounds, for example, when discussing news as further context for sonification. Clearly, some prefer “facts not thrills” (P6) and want their information not emotionalized.

Further, some users deliberately do not want that triggering of negative feelings. Therefore, designers might also aim to balance hazardous weather conditions like thunder with sounds that indicate a positive feeling of a safe place or home. Nonetheless, future studies could deeply focus on the relation between voice and (weather) sounds to experiment with fitting voice modulations that mirror the context. In general, sound bears an opportunity to reinforce calming situations, as raindrops against the window were positively associated.

6.3 Limitations

Our study investigated just one potential use case of sound overlays and VAs. However, a further holistic investigation is needed that requires testing several use cases to thoroughly understand how to use sound in voice interaction. Nevertheless, we could observe positive reactions to our design.

We mainly focused on developing a design approach and examining the general feasibility of a basic concept. At this point, we did not include advanced methods to examine the discriminative quality of the voice within our sonic overlays. Hence, we expect room for improvement in this area. In future work, additional quantitative studies, e.g., asking participants to transcribe the speech of the VA afterward, and using established Quality of Experience measurements as applied in telecommunications engineering [74], might significantly optimize the discriminative quality.

The same holds for our insights into semantic mapping and sonic associations. In the tradition of explorative qualitative research [75], our study uncovers relations and suggests hypotheses without statistical validation. For instance, our study suggests that the mapping and sonic associations are more coherent, when the illustrating situation (e.g., “it is raining”) refers to natural sounds. Future studies should evaluate our insights and implications quantitatively to gain validated results that either confirm our hypotheses or show further areas of improvement [75].

Furthermore, the examples we tested were not representing real-time weather conditions at the location of our participants, nor were they presented in a realistic situation, e.g., during time pressure or participants knowing they need to leave the house in the next 10 min. To provide more robust results, tests need to be investigated that resemble both more realistic situations and feature the actual outdoor weather situation. Finally, our test was based on a rudimentary prototype that was not implemented and run on an actual smart speaker. We think that rerunning our study in a realistic and practice-based context might reveal further design principles and limits of usability but also opportunities for more sonic design.

7 Conclusion

We presented a study that aims to investigate what designers can learn from sound design if they like to enrich the experience with Voice Assistants. Focusing on one of the most favorite use cases, we present a user-centered approach to designing sonic overlays that complement the vocal messages of Voice Assistants and contribute to its user experience. Specifically, we were interested in how sonification of data might enhance voice interaction by using iconic, associated, and abstract sounds, in the example of weather forecasts. Based on a pre-study with 48 participants, we constructed a sound library for creating soundscapes for seven weather conditions: sunny, cloudy, foggy, thunder, rainy, freezing, snowing. We further evaluated the resulting soundscapes in an interview study with 15 participants to learn more about the effects of underlying spoken information with complementing soundscapes. Our study revealed both positive and negative feedback from our interviewees, based on which we were able to elicit respective design implications. Our design approach aims to open the design space for further sonic investigations and designs enriching voice interaction.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. McLean G, Osei-Frimpong K (2019) Hey Alexa ... examine the variables influencing the use of artificial intelligent in-home voice assistants. *Comput Human Behav* 99(April):28–37. <https://doi.org/10.1016/j.chb.2019.05.009>

2. Porcheron M, Fischer JE, Reeves S, Sharples S: Voice Interfaces in Everyday Life. In: Proc. 2018 CHI Conf. Hum. Factors Comput. Syst. - CHI '18, vol. 2018-April, pp. 1–12. ACM Press, New York, New York, USA (2018). <https://doi.org/10.1145/3173574.3174214>
3. Bentley F, Luvogt C, Silverman M, Wirasinghe R, White B, Lottridge D: Understanding the Long-Term Use of Smart Speaker Assistants. Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol. 2(3), 1–24 (2018). <https://doi.org/10.1145/3264901>
4. Clark L, Doyle P, Garaialde D, Gilmartin E, Schlögl S, Edlund J, Aylett M, Cabral J, Munteanu C, Edwards J, R Cowan, B: The State of Speech in HCI: Trends, Themes and Challenges. Interact Comput 31(4), 349–371 (2019) 1810.06828. <https://doi.org/10.1093/iwci/iwz016>
5. Ammari T, Kaye J, Tsai JY, Bentley F: Music, Search, and IoT: How people (really) use voice assistants. ACM Trans Comput Interact 26(3) (2019). <https://doi.org/10.1145/3311956>
6. Sciuto A, Saini A, Forlizzi J, Hong JI: Hey Alexa, what's up?: Studies of in-home conversational agent usage. In: DIS 2018 - Proc. 2018 Des. Interact. Syst. Conf., pp. 857–868. ACM Press, New York, New York, USA (2018). <https://doi.org/10.1145/3196709.3196772>
7. Akçay MB, Oğuz K (2020) Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication 116:56–76
8. Koolagudi SG, Rao KS (2012) Emotion recognition from speech: a review. Int J Speech Technol 15(2):99–117
9. Schuller B, Rigoll G, Lang M: Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 1, p 577. IEEE, USA (2004)
10. Schröder M: Emotional speech synthesis: A review. In: Seventh European Conference on Speech Communication and Technology. Citeseer, Aalborg, Denmark (2001)
11. Schmitt A, Zierau N, Janson A, Leimeister JM: Voice as a contemporary frontier of interaction design. In: European Conference on Information Systems (ECIS).-Virtual (2021)
12. Demaeght A, Nerb J, Müller A: A survey-based study to identify user annoyances of german voice assistant users. In: International Conference on Human-Computer Interaction, pp 261–271. Springer, Cham (2022)
13. Cabral JP, Cowan BR, Zibrek K, McDonnell R: The Influence of Synthetic Voice on the Evaluation of a Virtual Character. In: Proc. Interspeech 2017, pp 229–233 (2017). <https://doi.org/10.21437/Interspeech.2017-325>
14. Weiss B, Trouvain J, Barkat-Defradas M, Ohala JJ (2021) Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers. Springer, Singapore
15. Sutton SJ: Gender ambiguous, not genderless: Designing gender in voice user interfaces (vuis) with sensitivity. In: Proceedings of the 2nd Conference on Conversational User Interfaces, pp 1–8 (2020)
16. Kilian K, Kreutzer RT: In: Kilian K, Kreutzer RT (eds.) Voice-Marketing, pp 279–312. Springer, Wiesbaden (2022). https://doi.org/10.1007/978-3-658-34351-4_9
17. Klein AM, Hinderks A, Schrepp M, Thomaschewski J: Measuring user experience quality of voice assistants. In: 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), pp 1–4. IEEE, Danvers, MA (2020)
18. Cho M, Lee S-s, Lee K-p: Once a Kind Friend is Now a Thing. In: Proc. 2019 Des. Interact. Syst. Conf. - DIS '19, pp 1557–1569. ACM Press, New York, New York, USA (2019). <https://doi.org/10.1145/3322276.3322332>
19. Parviainen E, Søndergaard MLJ: Experiential Qualities of Whispering with Voice Assistants. Conf. Hum. Factors Comput. Syst. - Proc., 1–13 (2020). <https://doi.org/10.1145/3313831.3376187>
20. Simpson J: Are CUIs Just GUIs with Speech Bubbles? In: Proc. 2nd Conf. Conversational User Interfaces, pp 1–3. ACM, New York, NY, USA (2020). <https://doi.org/10.1145/3405755.3406143>
21. Chavez-Sanchez F, Franco GAM, de la Peña GAM, Carrillo EIH: Beyond What is Said. In: Proc. 2nd Conf. Conversational User Interfaces, pp 1–3. ACM, New York, NY, USA (2020). <https://doi.org/10.1145/3405755.3406145>
22. Rönneberg N: Sonification for Conveying Data and Emotion. In: Audio Most. 2021, pp 56–63. ACM, New York, NY, USA (2021). <https://doi.org/10.1145/3478384.3478387>
23. Juslin PN: What does music express? Basic emotions and beyond. Front. Psychol. 4(SEP) (2013). <https://doi.org/10.3389/fpsyg.2013.00596>
24. Blattner M, Sumikawa D, Greenberg R (1989) Earcons and Icons: Their Structure and Common Design Principles. Human-Computer Interact. 4(1):11–44. https://doi.org/10.1207/s15327051hci0401_1
25. Gaver WW (1989) The SonicFinder: An Interface That Uses Auditory Icons. Human-Computer Interact. 4(1):67–94. https://doi.org/10.1207/s15327051hci0401_3
26. Liljedahl M, Fagerlönn J: Methods for sound design. In: Proc. 5th Audio Most. Conf. A Conf. Interact. with Sound - AM '10, pp 1–8. ACM Press, New York, New York, USA (2010). <https://doi.org/10.1145/1859799.1859801>
27. Nass C, Lee KM: Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp 329–336 (2000)
28. Sutton SJ, Foulkes P, Kirk D, Lawson S: Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp 1–14 (2019)
29. Frederking RE (1996) Grice's maxims: do the right thing. Frederking, RE
30. Esau M, Krauß V, Lawo D, Stevens G: Losing its touch: Understanding user perception of multimodal interaction and smart assistance. In: Designing Interactive Systems Conference. DIS '22, pp 1288–1299. Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3532106.3533455>
31. Myers C, Furqan A, Nebolsky J, Caro K, Zhu J: Patterns for how users overcome obstacles in Voice User Interfaces. In: Conf. Hum. Factors Comput. Syst. - Proc., vol. 2018-April, pp 1–7. ACM Press, New York, New York, USA (2018). <https://doi.org/10.1145/3173574.3173580>
32. Cowan BR, Pantidi N, Coyle D, Morrissey K, Clarke P, Al-Shehri S, Earley D, Bandeira N: what can i help you with? infrequent users' experiences of intelligent personal assistants. In: Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp 1–12 (2017)
33. Burmester M, Zeiner K, Schippert K, Platz A: Creating positive experiences with digital companions. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pp 1–6 (2019)
34. Aylett MP, Clark L, Cowan BR: Siri, echo and performance: You have to suffer darling. In: Conf. Hum. Factors Comput. Syst. - Proc., pp 1–10. ACM Press, New York, New York, USA (2019). <https://doi.org/10.1145/3290607.3310422>
35. Hassenzahl M, Borchers J, Boll S, der Pütten AR-v, Wulf V: Otherware: how to best interact with autonomous systems. Interactions 28(1), 54–57 (2021). <https://doi.org/10.1145/3436942>
36. Luger E, Sellen A: Like having a really bad pa: The gulf between user expectation and experience of conversational agents. In: Conf. Hum. Factors Comput. Syst. - Proc., pp 5286–5297. ACM Press, New York, New York, USA (2016). <https://doi.org/10.1145/2858036.2858288>

37. Petty RE, Brinol P: The Elaboration Likelihood Model, pp 224–245. SAGE, London, UK (2011)
38. Petty RE, Cacioppo JT: Source factors and the elaboration likelihood model of persuasion. *ACR North American Advances NA-11* (1984)
39. Lugovic S, Dunder I, Horvat M: Techniques and applications of emotion recognition in speech. In: 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (mipro), pp 1278–1283. IEEE, Rijeka, Croatia (2016)
40. Lee C-C, Kim J, Metallinou A, Busso C, Lee S, Narayanan SS: Speech in affective computing, pp 170–183. Oxford Univ. Press New York, NY, USA, New York, USA (2014)
41. Juslin PN, Laukka P (2003) Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin* 129(5):770
42. Seaborn K, Miyake NP, Pennefather P, Otake-Matsuura M (2021) Voice in human-agent interaction: a survey. *ACM Computing Surveys (CSUR)* 54(4):1–43
43. Schuller B, Batliner A, Steidl S, Seppi D (2011) Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech communication* 53(9–10):1062–1087
44. Burkhardt F, Stegmann J (2009) Emotional speech synthesis: Applications, history and possible future. *Studentexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung* 2009:190–199
45. Eide E, Aaron A, Bakis R, Hamza W, Picheny M, Pitrelli J: A corpus-based approach to expressive speech synthesis. In: Fifth ISCA Workshop on Speech Synthesis (2004)
46. Shi Y, Yan X, Ma X, Lou Y, Cao N: Designing emotional expressions of conversational states for voice assistants: Modality and engagement. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, pp 1–6 (2018)
47. Salselas I, Penha R, Bernardes G (2021) Sound design inducing attention in the context of audiovisual immersive environments. *Pers. Ubiquitous Comput.* 25(4):737–748. <https://doi.org/10.1007/s00779-020-01386-3>
48. Enge K, Rind A, Iber M, Höldrich R, Aigner W: It's about Time: Adopting Theoretical Constructs from Visualization for Sonification. In: *Audio Most.* 2021, pp 64–71. ACM, New York, NY, USA (2021). <https://doi.org/10.1145/3478384.3478415>
49. Mansur DL, Blattner MM, Joy KI: Sound graphs: A numerical data analysis method for the blind. *Journal of medical systems* 9(3), 163–174 (1985)
50. Nesbitt KV, Barrass S: Evaluation of a multimodal sonification and visualisation of depth of market stock data. (2002). Georgia Institute of Technology
51. Brewster SA: Providing a structured method for integrating non-speech audio into human-computer interfaces (1994)
52. Seiça M, Roque L, Martins P, Cardoso FA: Contrasts and similarities between two audio research communities in evaluating auditory artefacts. In: Proc. 15th Int. Conf. Audio Most., pp 183–190. ACM, New York, NY, USA (2020). <https://doi.org/10.1145/3411109.3411146>
53. Campbell IG (1942) Basal emotional patterns expressible in music. *The American Journal of Psychology* 55(1):1–17
54. Donald R, Kreutz G, Mitchell L, MacDonald R: What is music health and wellbeing and why is it important? In: *Music, Health, and Wellbeing*, pp 3–11. Oxford University Press, Oxford (2012)
55. Juslin PN, Sloboda J (2011) *Handbook of Music and Emotion: Theory, Research. Applications.* Oxford University Press, Oxford
56. Jerald J (2015) *The VR Book: Human-centered Design for Virtual Reality.* Morgan & Claypool, New York, USA
57. Carvalho FR, Steenhaut K, van Ee R, Touhafi A, Velasco C: Sound-enhanced gustatory experiences and technology. In: Proc. 1st Work. Multi-sensorial Approaches to Human-Food Interact. MHFI '16, pp 1–8. ACM, New York, NY, USA (2016). <https://doi.org/10.1145/3007577.3007580>
58. Wang QJ, Mesz B, Spence C: Assessing the impact of music on basic taste perception using time intensity analysis. In: Proc. 2nd ACM SIGCHI Int. Work. Multisensory Approaches to Human-Food Interact. MHFI 2017, pp 18–22. ACM, New York, NY, USA (2017). <https://doi.org/10.1145/3141788.3141792>
59. ISO Norm: Acoustics – Soundscape – Part 1: Definition and conceptual framework. Last accessed 31.03.2022 (2022). <https://www.iso.org/standard/52161.html> Accessed 13 Mar 2022
60. Hong J, Yi HB, Pyun J, Lee W: SoundWear: Effect of non-speech sound augmentation on the outdoor play experience of children. In: DIS 2020 - Proc. 2020 ACM Des. Interact. Syst. Conf., pp 2201–2213. ACM, New York, NY, USA (2020). <https://doi.org/10.1145/3357236.3395541>
61. Chung D, Tsai W-C, Liang R-H, Kong B, Huang Y, Chang F-C, Liu M: Designing Auditory Experiences for Technology Imagination. In: 32nd Aust. Conf. Human-Computer Interact., pp 682–686. ACM, New York, NY, USA (2020). <https://doi.org/10.1145/3441000.3441025>
62. Mynatt ED: Designing with Auditory Icons (1994)
63. Krygier JB: Sound and geographic visualization. In: *Mod. Cartogr. Ser. vol 2*, pp 149–166. Elsevier Science Ltd, Kidlington, Oxford, OX5 1GB, U.K. (1994). <https://doi.org/10.1016/B978-0-08-042415-6.50015-6>
64. Peirce CS (1991) *Peirce on Signs: Writings on Semiotic.* UNC Press Books, North Carolina
65. Oswald D: Non-speech audio-semiotics: a review and revision of auditory icon and earcon theory (2012)
66. Whittington WB (1999) *Sound Design and Science Fiction.* University of Southern California, New York, USA
67. McGookin D, Brewster S: *Earcons.* Logos Publishing House, Berlin, Germany. (2011). publisher: Logos Verlag
68. Cabral JP, Remijn GB (2019) Auditory icons: Design and physical characteristics. *Appl Ergon* 78(January):224–239. <https://doi.org/10.1016/j.apergo.2019.02.008>
69. Biernacki P, Waldorf D (1981) Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods & research* 10(2):141–163
70. Guest G, MacQueen KM, Namey EE (2011) *Applied Thematic Analysis.* SAGE Publications, Los Angeles, USA
71. Clark L, Munteanu C, Wade V, Cowan BR, Pantidi N, Cooney O, Doyle P, Garaialde D, Edwards J, Spillane B, Gilmartin E, Murad C: What Makes a Good Conversation? In: Proc. 2019 CHI Conf. Hum. Factors Comput. Syst. - CHI '19, pp 1–12. ACM Press, New York, New York, USA (2019). <https://doi.org/10.1145/3290605.3300705>
72. Chion M (1994) *Audio-vision: Sound on Screen.* Columbia University Press, News York, USA
73. Yewdall DL (2012) *Practical Art of Motion Picture Sound.* Taylor & Francis, Waltham, MA, USA
74. Streijl RC, Winkler S, Hands DS: Mean opinion score (mos) revisited: methods and applications, limitations and alternatives 22, 213–227 (2016). <https://doi.org/10.1007/s00530-014-0446-1>
75. Kelle U, Erzberger C: *Qualitative and quantitative methods: not in opposition*, pp 172–177. SAGE Publications, London (2004). publisher: Sage Publications London

76. Lee S, Lee S, Kim S: What does your agent look like? A drawing study to understand users' perceived persona of conversational agent. In: Conf. Hum. Factors Comput. Syst. - Proc., pp 1–6. ACM Press, New York, New York, USA (2019). <https://doi.org/10.1145/3290607.3312796>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.