SoundsRide: Affordance-Synchronized Music Mixing for In-Car Audio Augmented Reality

Mohamed Kari
Porsche AG
University of Duisburg-Essen

Tobias Grosse-Puppendahl
Porsche AG

Alexander Jagaciak
Porsche AG

David Bethge
Porsche AG
LMU Munich

Reinhard Schütte
University of Duisburg-Essen

Christian Holz
Department of Computer Science
ETH Zürich

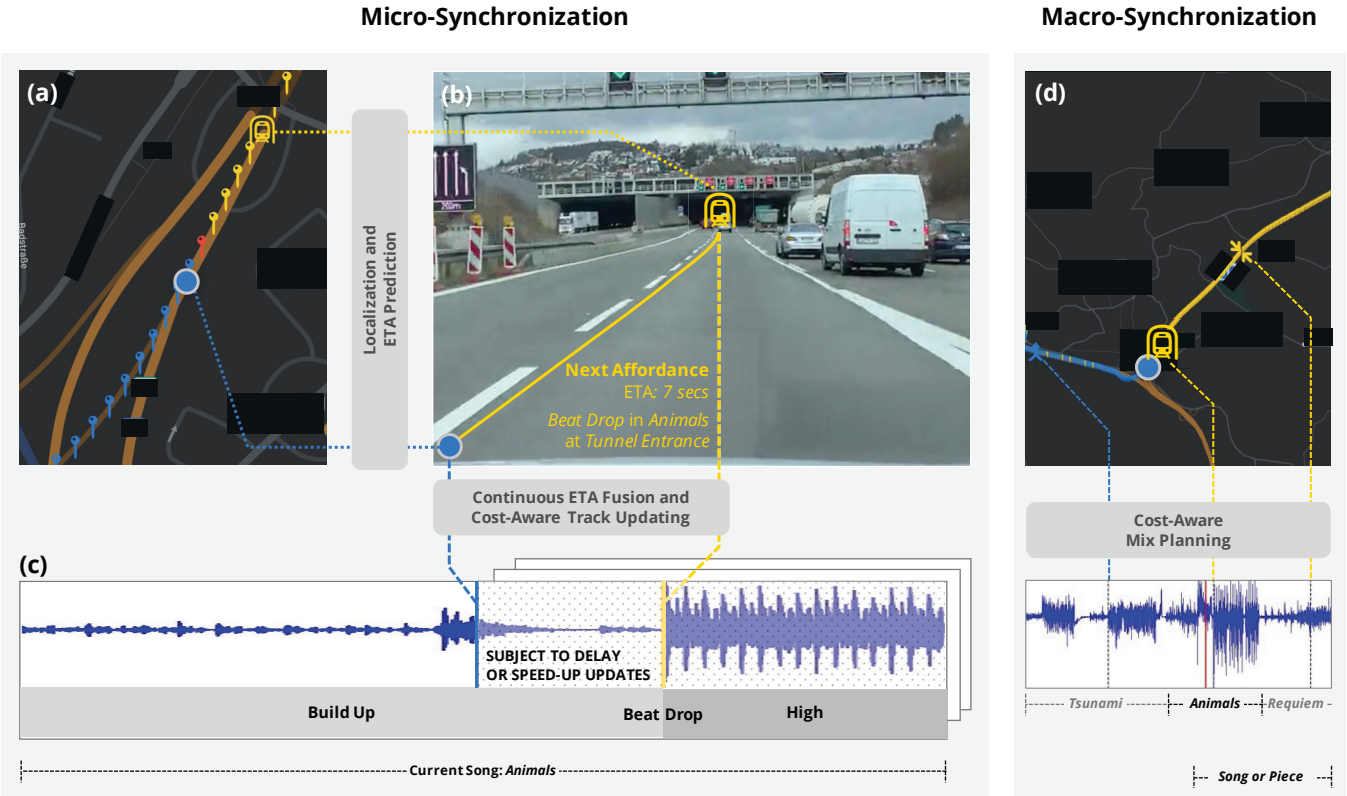


Figure 1: SoundsRide is an in-car audio augmented reality system that mixes music in real-time synchronized with sound affordances along the ride. (a) SoundsRide continuously predicts the Estimated Time to Arrival (ETA) to the next affordances (b) to temporally and spatially align high-contrast events on the ride – such as a tunnel entrance – (c) with high-contrast events in the music – such as a beat drop – by speeding up or delaying a mix event through track updates if necessary. (d) On a macro-level, SoundsRide makes use of affordance ETAs to plan a cost-aware mix for the entire ride by deciding on the song sequence and when to transition between songs.

UIST '21, October 10–14, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8635-7/21/10…$15.00
https://doi.org/10.1145/3472749.3474739

ABSTRACT

Music is a central instrument in video gaming to attune a player's attention to the current atmosphere and increase their immersion in the game. We transfer the idea of scene-adaptive music to car drives and propose SoundsRide, an in-car audio augmented reality system that mixes music in real-time synchronized with sound affordances along the ride. After exploring the design space of

affordance-synchronized music, we design SoundsRide to temporally and spatially align high-contrast events on the route, e. g., highway entrances or tunnel exits, with high-contrast events in music, e. g., song transitions or beat drops, for any recorded and annotated GPS trajectory by a three-step procedure. In real-time, SoundsRide 1) estimates temporal distances to events on the route, 2) fuses these novel estimates with previous estimates in a cost-aware music-mixing plan, and 3) if necessary, re-computes an updated mix to be propagated to the audio output. To minimize user-noticeable updates to the mix, SoundsRide fuses new distance information with a filtering procedure that chooses the best updating strategy given the last music-mixing plan, the novel distance estimations, and the system parameterization. We technically evaluate SoundsRide and conduct a user evaluation with 8 participants to gain insights into how users perceive SoundsRide in terms of mixing, affordances, and synchronicity. We find that SoundsRide can create captivating music experiences and positively as well as negatively influence subjectively perceived driving safety, depending on the mix and user.

CCS CONCEPTS

• **Human-centered computing → Mixed / augmented reality**; **Ubiquitous and mobile computing systems and tools**.

KEYWORDS

auditory augmented reality, sound affordances, context-adaptive music

ACM Reference Format:
Mohamed Kari, Tobias Grosse-Puppendahl, Alexander Jagaciak, David Bethge, Reinhard Schütte, and Christian Holz. 2021. SoundsRide: Affordance-Synchronized Music Mixing for In-Car Audio Augmented Reality. In *The 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21), October 10–14, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3472749.3474739

1 INTRODUCTION

Music has become an integral part of a driver's in-car experience. It is powerful in influencing the driver's mental state, being able to provoke negative as well as positive effects on driving experience and performance through mechanisms of arousal and distraction or concentration in the driver [5, 12, 26, 34, 37, 40]. More pragmatically, music has been instrumented purposefully to assist in tasks such as keeping a certain speed [7] or navigating towards a specified destination [1].

To enhance a user's music experience, previous work has explored adapting music to the driver's context based on their internal state, behavior, or environment. However, music adaption to the user's internal state [10] is challenging due to the complex nature of mental state and the difficulties in estimating and influencing it purposefully. Approaches that adapt music to a user's behavior [11, 23] do not capture the potential to increase situational awareness for the driving task. Previous approaches that adapt music to a user's environment typically serve as mere decision support systems [2, 4, 8, 18, 32] without enabling novel forms of user experiences.

In this paper, we propose *SoundsRide*, an in-car audio augmented reality system that mixes music in real-time synchronized with the events surrounding the driver's ride. To categorize such events, we introduce a series of *sound affordances* that describe such momentous and well-noticeable events in the environment. For example, the exit of a tunnel with its stark switch in illumination and acoustics affords a strong musical event such as a beat drop or a crescendo that builds up until the instant of the vehicle exiting the tunnel. We argue that aligning such high-contrast events along the ride with high-contrast events in the music can create captivating experiences shaped by memorable moments along the ride and not yet achievable by users on their own without such technical enablement.

The core technical contribution of SoundsRide lies in its capability of synchronizing music to affordances in real-time by creating a mix from a song database so that high-contrast intra-song or inter-song events in the mixed audio signal are temporally and spatially aligned with the sound affordances. Precisely scheduling such auditory signals in response to a driving car is challenging and comprises three subproblems: 1) localizing sound affordances on the route and predicting an estimated time to arrival ("affordance ETAs"), 2) determining a mixing plan that minimizes undesirable effects such as silences between songs, and 3) continuously integrating updated affordance ETAs without causing disruptions in the mixed audio signal.

In the approach section, we first describe the design alternatives of affordance-synchronized music and then detail SoundsRide's integration of geo-referenced affordances to enable ETA predictions, our heuristics-based scheduling method for affordance-synchronized music mixes, as well as our recursive filtering technique for updating the audio signal to updated ETA predictions.

In our qualitative evaluation, eight participants drove a route with SoundsRide and reported on their impression of the ride. Participants generally commented positively on the drive and seven pointed out their excitement of experiencing music that adapts to the scene. The results also showed that SoundsRide helped participants become more aware of their environment while driving, as they freely listed half of all affordances without explicitly having been instructed about them; they remembered 18 % more affordances when explicitly asked about them, and recognized another quarter of affordances when watching a video replay of their ride.

In our quantitative evaluation of SoundsRide's performance, we examined the level of affordance synchronicity achieved based on the rides recorded during the participant evaluation. SoundsRide ensured synchronicity within $\pm 1.1\,s$ in 47.7 % of cases and with no more than 1 noticeable update to the audio signal within 15 seconds of an affordance location. In 77.7 % of cases, SoundsRide is able to ensure synchronicity with a maximum misalignment of $1.9s$ and a maximum number of 2 updates to the audio track.

Contributions

In this paper, we make the following contributions:

(1) a novel end-to-end approach for creating suspenseful and affordance-synchronized in-car music experiences based on temporal distance estimation for reinforcing high-contrast events along a ride with high-contrast events in music,

(2) a design space of affordance-based in-car audio augmented reality for music experiences,
(3) a cost-based method for deriving affordance-synchronized mix plans featuring inter- and intra-song mix events with song snippets from an annotated song databases, and
(4) a recursive filter algorithm for incorporating continuously updated ETA information while minimizing obtrusive audio track manipulations.

Combining these contributions in a real-time system, SoundsRide brings dynamic and environment-aware audio augmented reality to everyday car rides, supporting the joy of driving and drawing the driver's attention to the periphery. We provide our implementation of SoundsRide and accompanying assets to the community for future work[1].

2 BACKGROUND AND RELATED WORK

SoundsRide's idea of affordance-synchronized music mixing builds on concepts found in context-adaptive music, music-to-video alignment, procedural game music, and in-transit audio augmented reality.

2.1 Context-Adaptive Music

Approaches that aim to adapt music to a user's internal state draw on user-focusing context variables such as emotion, mood, or fatigue [2, 10, 15, 18, 39]. These context-adaptive approaches often operate on difficult-to-test assumptions about the complex mental dynamics and interdependent processes in humans that govern music listening preferences in a specific real-world situation.

Within the field of approaches that aim to adapt music to a user's environment, location-aware music recommendation [2, 4, 8, 16, 17, 23, 32] takes a particularly prominent role. However, the main objective of these music recommender systems typically consists of improving or simplifying a user's *decision making* in terms of music selection by drawing on typically coarse-grained location classes as decision parameters.

Approaches that aim to adapt music to a user's externally observable behavior, such as pace [11] or driving style [2] reflect and might therefore reinforce a user's activity, however do not expose interactivity with the environment around the user.

In contrast to location-aware music recommendation systems, SoundsRide does not aim at supporting decision making but rather at enabling scene-synchronized music mixing based on accurate estimation of temporal distance to sound affordance moments for captivating and suspenseful experiences. In contrast to behavior-adaptive music playback, SoundsRide is bound to sound affordances in the environment, therefore featuring a novel user-environment interaction pattern. In contrast to affect-adaptive music playback, beyond the fundamental hypothesis that scene-synchronized music is interesting to users, SoundsRide does not make implicit assumptions about the complex interrelationship between affective state and music.

[1] https://github.com/MohamedKari/soundsride

2.2 Automatic Music-to-Video Alignment

UnderScore by Rubin et al. [30] automatically derives a musical underlay for an audio story, constrained by emphasis points in speech. Rubin et al. [29] build on the idea of UnderScore, however focus on emotions as a key constraint under which alignment takes place. Sato et al. [31] present a system that automatically arranges a soundtrack so that it fits climaxes in a video. Wang et al. [36] present a system that synthesizes background music after visually classifying intervals in a given video by emotions. Frid et al. [13] propose a system that allows MIDI-based synthesis of music, similar to a provided reference song, in order to acoustically underlay a video.

While these systems are not location-aware, they are abstractly similar in that they align music with externally specified moments in time. However, they are very different in that they operate on all video data available, while SoundsRide operates in real-time and under uncertainty, thus requiring not only a prediction but also a correction procedure for incorporating updated affordance ETAs into the signal output.

2.3 Procedural Game Music

To increase immersion [6, 28], many games feature a rich soundscape, ranging from simple player-controlled sounds such as footsteps or collecting coins to atmospheric and narrative-supporting game scores that are bound to certain trigger points in the game's visual space. Creating these soundscapes is often based on so-called procedural or non-linear music composition that allows to add in, subtract, transpose, or swap layers of instrumentation or parameterize a predefined musical sequence in terms of jumps, repeats, or loops [9, 38]. The landmark system iMuse by Land and McConnell [22] from 1991 has heralded non-linear music in gaming by enabling seamless transitions, triggered by switching the gameplay levels or certain events in the gameplay and based on decision points in the score that allow branching to one or the other sequence [27].

SoundsRide's problem statement is similar to such game score engines as it temporally synchronizes music with trigger points, however is different in that it aligns *intra-song events in common songs* with these trigger points rather than branching from a dedicated score. As a consequence, to schedule a song ahead, SoundsRide predicts the time to arrival to the next trigger point, which due to the fewer degrees of freedom in driving, can be more deterministically planned than a player's gaming interactions.

2.4 In-Transit Audio AR

We refer to the common notion of Audio AR as superimposing an audio signal on top of the real world, as the user moves within it [3, 21, 24]. On a basic level, common navigation systems with speech output such as Google Maps[2] or Waze[3] can be interpreted as audio augmented reality applications that supply context-bound information to drivers, pedestrians, or cyclists on the auditory channel. Systems like GyPSy Guide[4] provide audio commentary playback for in-car usage along a path of specified locations to provide a location-bound virtual tour guide.

[2] https://maps.google.com/
[3] https://waze.com/
[4] https://gypsyguide.com/

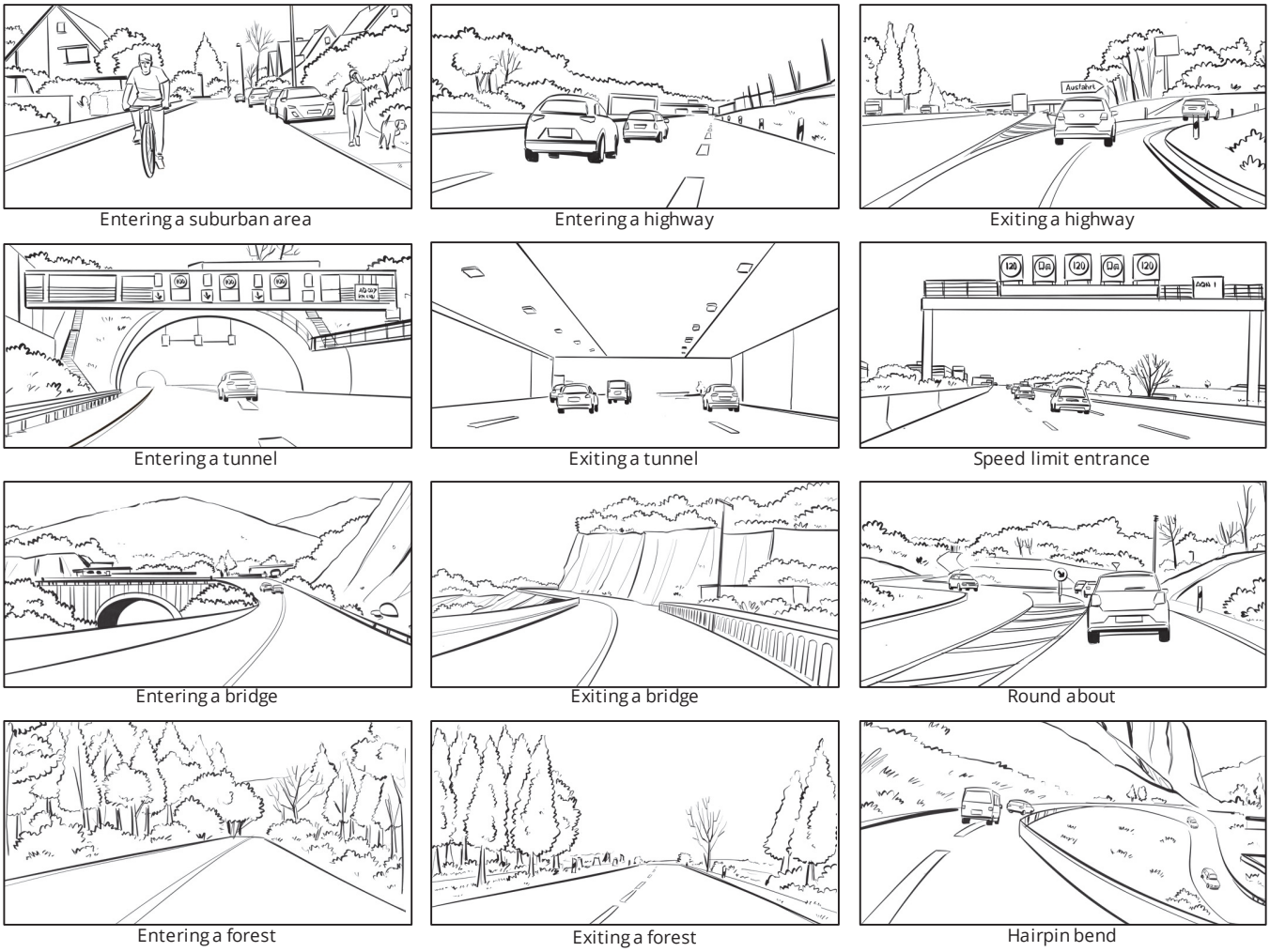


Figure 2: Examples for location-bound affordance situations. SoundsRide's design allows annotating any of these affordance situations on the route and to temporally and spatially align events in the music to them.

On a more sophisticated level, HindSight [33] employs a sonification of objects detected in real time in continuously streamed 360° video to increase an in-transit user's awareness of vehicles in the surroundings. Audible Panorama by Huang et al. [19] augments a panorama picture with an audio signal, generated by first detecting objects in the image and then mixing sounds corresponding to the object class.

Both aforementioned systems are conceptually distinct from the music-to-video alignment task because they take a much more high-resolution understanding of reality as a basis – using machine-learned computer vision algorithms – in contrast to operating on a series of non-further qualified video frames. Even though Audible Panorama is not focused on in-transit usage, both the ideas in HindSight and Audible Panorama can be interpreted as "affordance-oriented" in that they assign certain artificial sounds to real-world observations.

However, SoundsRide is different from both these systems in that 1) it focuses on music instead non-musical sounds, 2) must plan ahead and update plans under location-awareness in real-time to temporally align mix events with environment events, and 3) opens up interesting user-controlled interaction patterns between the user, the system, and the environment.

3 AFFORDANCE-BASED MUSIC

3.1 Sound Affordances: Features in the environment that lend themselves to acoustic events

Affordances are action possibilities [20]. We define *sound affordances* as momentous and well-noticeable events in the environment, characterized by a salient contrast in some user-perceivable aspect and offering the possibility to assign a certain musical event. These contrastive aspects are not limited to visual contrasts – such

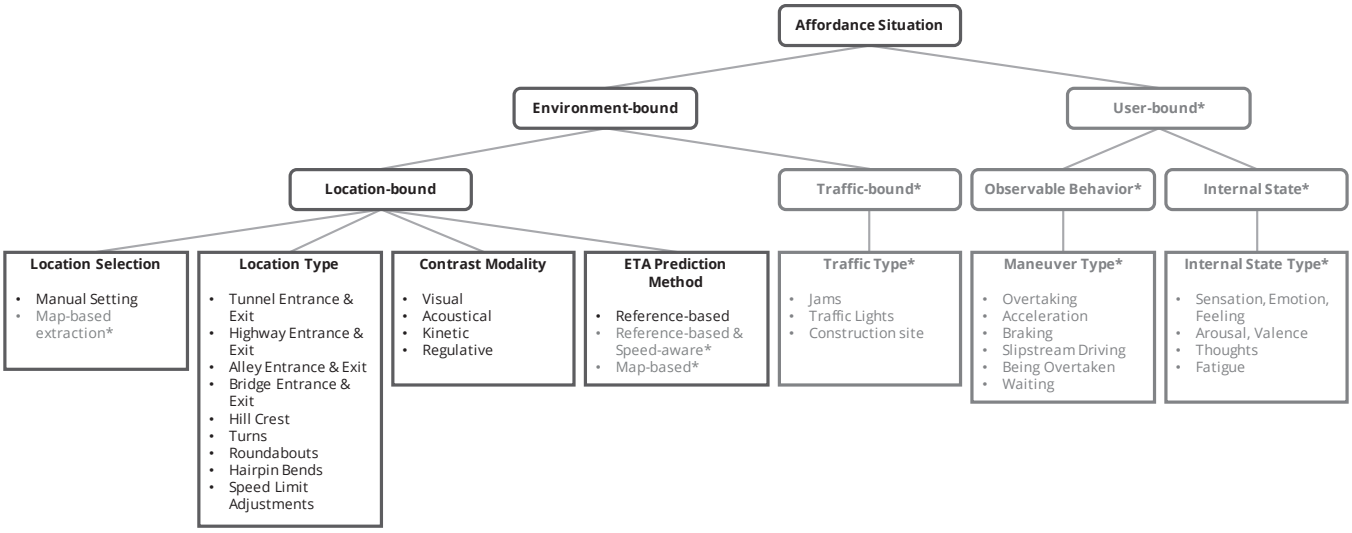


Figure 3: Taxonomy of affordance situations in affordance-based music. Design choices for SoundsRide are shown in black (without asterisk). SoundsRide is focused on location-bound affordance situations that are annotated by the user with a mobile app on a self-recorded GPS trajectory. This enables users to label any location as an affordance situation of a certain type and underline any contrast modality with music. The same GPS trajectory used for annotation is also used for affordance ETA prediction.

as entering or exiting a tree-lined alley, but might alternatively or additively refer to a contrast in *g*-force – e. g., when passing the crest of a hill, a contrast in sound perception – e. g., when entering or exiting a tunnel, a contrast in the traffic guidances – e. g., specification or revocation of a speed limit, or a contrast in the overall traffic flow – e. g., when entering or exiting a highway. In our understanding, sound affordances are constituted by an *affordance situation* that provides the opportunity for an *affordance action*.

Sound affordance situations can be either bound to the environment or bound to the user. User-bound affordances are endogenous and can either focus on the users' internal state or their observable behavior. On the other hand, environment-bound affordances are exogenous and comprise locations and traffic situations. While variables such as weather or daytime can serve as more general context variables, they do not represent contrastive events in a ride that can be annexed for events in music. Within the space of location-bound affordance situations, relevant dimensions for positioning a location-bound affordance situation include the type of the location, the modality under which the contrast unfolds, the way it can be determined as an affordance, and the method that can be used to estimate the temporal distance to it. Figure 2 shows a set of examples for location-bound affordance situations.

A *sound affordance action* is taken by the affordance-handling system to create a musical audio signal from information about affordance situations. Context-aware music recommender systems such as [16] or [17] aim to reflect the *mood* in a driving segment with a length of minutes or tens of minutes in the music. We call this objective *macro-synchronicity*. Alignment in affordance-based music, first and foremost, aims at *micro-synchronicity*. Micro-synchronicity is characterized by temporal alignment of *events* in the music mix and estimated events in the environment in a subsecond-to-seconds time horizon. Temporal alignment will ensure spatial alignment given accurate ETA predictions. While micro-synchronicity can entail macro-synchronicity with a suitable song selection per affordance, the inverse is generally not true. Conceptually, concerning sound affordance actions, we distinguish between music playback, music mixing, and music generation. These different modes of affordance actions differ in terms of the degrees of freedom they can control to align music with affordance situations. In *music playback*, the song sequence is the only decision variable. This mode of operation is employed in context-aware music recommender systems and can only aim at macro-synchronicity. In contrast, music generation and music mixing can aim at micro-synchronicity. In *music generation*, MIDI-based techniques as found in procedural game music or music live-coding techniques are used to arrange elementary sounds or more complex instrument loops to create a novel piece of music, thus offering the maximum degree of freedom. In *music mixing*, decision variables include song-structure-aware fade-in and fade-out of songs, deliberate repetition of segments such as beats, bars, or parts, sound effects, frequency equalization or filtering, stretching, panning and balancing, transition effects, looping, etc. To ensure micro-synchronicity, a system needs to react to unexpected changes in ETA predictions by delaying or speeding-up a planned event in the mix using a *resynchronization strategy* such as BPM stretching or track resetting. Events in the mix to be aligned are *inter-song transitions* or *intra-song features* that are either recognized automatically or annotated manually.

3.2 Affordance-Synchronized Music Mixing in SoundsRide

Situation-wise, location-bound affordances are particularly interesting for SoundsRide as we assume that 1) they are intuitively

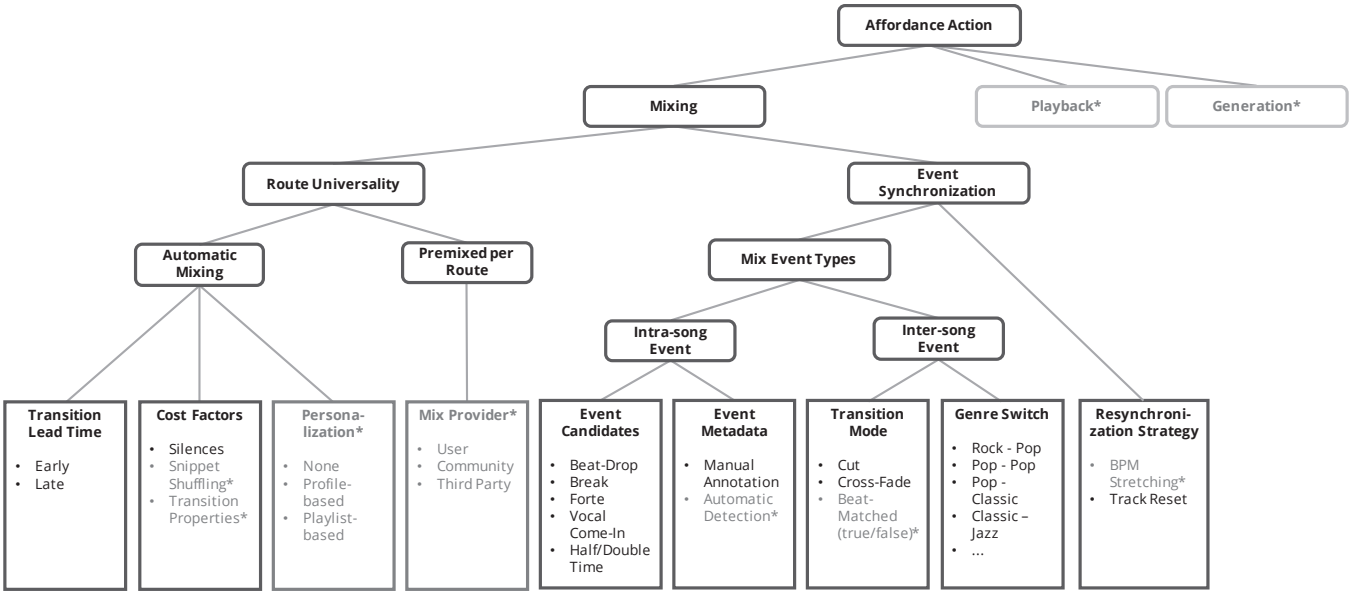


Figure 4: Taxonomy of affordance actions in affordance-based music. Design choices for SoundsRide are shown in black (without asterisk). SoundsRide is focused on mixing music so that intra-song or inter-song events are temporally and spatially synchronized with affordance situations. Synchronization takes place on a micro-level by track resetting alignment as well as on a macro-level by affordance-oriented song selection. It allows automatic mixing for any user-recorded and user-annotated GPS trajectory as well as aligning a provided route-specific mix for the current ride.

comprehensible to users, 2) are more robustly measurable than internal state, 3) at least partially capture the spirit of user-bound affordances implicitly, assuming that location influences the user's behavior and internal state, and 4) might increase the situational awareness beneficially for driving as opposed to user-bound affordances. Figure 3 summarizes our taxonomy of affordance situations and the position of SoundsRide.

Action-wise, for SoundsRide, we choose a music mixing approach that allows automatic mixing as well as aligning provided route-specific mixes for the current ride. We design SoundsRide for intra-song as well as inter-song events and use track resetting for delaying as well as speeding up planned events. Figure 4 summarizes our taxonomy of affordances and the position of SoundsRide.

In order to mix music based on location-bound sound affordances, SoundsRide tackles three problems: 1) Affordance ETA Prediction, 2) Mix Planning, and 3) Innovative Information Fusion.

3.2.1 Affordance ETA Prediction. The problem of estimating the temporal distance to a certain location in the route ahead is commonly solved in navigation systems such as Google Maps or Waze. However, the navigational use case of these systems typically tolerates a deviation of a couple of seconds. More precisely, navigational use cases do not require an accuracy exact to the second quite a time ahead of the actual event. However, to begin a music snippet so that an event in the music tens of seconds later is aligned with an upcoming sound affordance requires such an estimation - at least if last-second manipulations are to be avoided. We call this special case *micro-temporal estimation*.

For the purposes of our application, we approach the problem of micro-temporal estimation by allowing the user to first record a reference GPS trajectory, i. e., a path of GPS points. After recording, the user can mark points as affordance locations by tapping the respective pin (see Figure 1) and toggling through the offered affordance types. For the actual SoundsRide session, the user is re-localized against the reference GPS trajectory and the estimated time to arrival (ETA) to each of the succeeding sound affordances is easily computed as an aggregate over the temporal distance between pins. Figure 2 shows types of affordance situations that users might annotate on a route of their choosing, e. g., their commute.

3.2.2 Mix Planning. Given a specification of affordance ETAs from the step above, SoundsRide creates a mix plan with scheduled mix events. Figure 5 shows how a mix plan is generated from the affordance ETAs by aligning events in a song snippet (i. e., intra-song events) or transitions between songs (i. e., an inter-song event) with the ETAs, then transitioning between song snippets, and finally mixing the audio signal to be written to the played-back audio buffer. The mix event type is determined by a predefined mapping from affordance situations to mix event types. SoundsRide looks up songs in a database of annotated songs that contain events of the needed event type. From all matching song events, SoundsRide chooses the song that incurs the least cost. Cost is incurred if a song is so short that it ends before the next song can fade-in, hence resulting in silence. Overlaps between songs are eliminated by applying a set of cross-fading rules, making sure that cross-fades only take place beyond a safe zone around the aligned mix event. This step

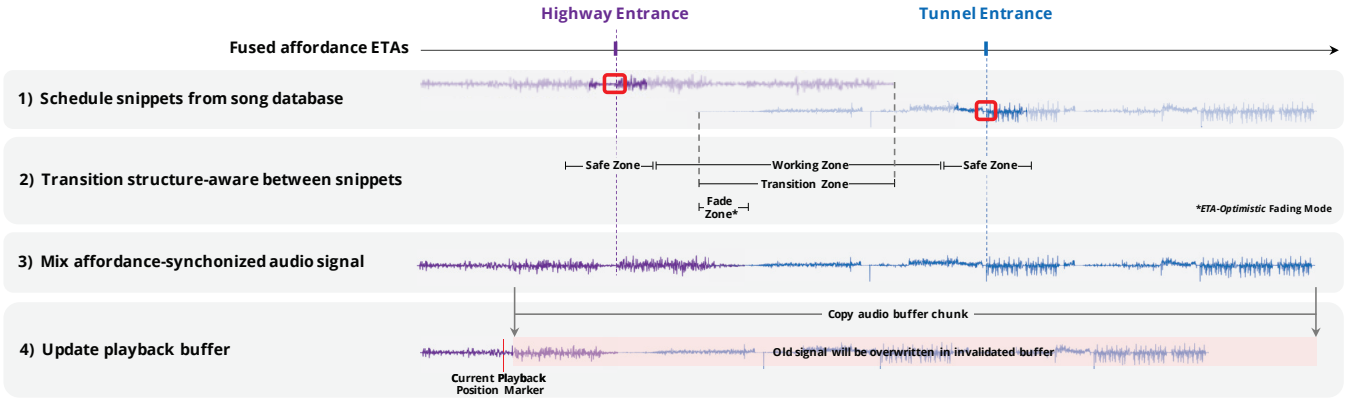


Figure 5: Whenever a resynchronization is triggered by the information fusion, SoundsRides overwrites the played-back audio buffer with an updated audio signal. This audio signal is produced by 1) aligning events in or transitions between songs snippets from a song database on an audio track along a given set of affordance ETAs, 2) fading these songs snippets in at some point in time, and fading them out at some point later in time. Fading times are based on the annotated structure of the song and the overlaps between the snippets. 3) ETA-aware positioning ensures alignment of song events in the mix with affordances in the environment. 4) Future chunks of the played-back audio buffer are then updated in-place.

features a configurable parameter *ETA accuracy optimism* that determines how early or late the next song is faded in before the next affordance situation. Fading in early means that non-negligible updates to the next affordance ETA will become more likely, whereas fading in late, e. g., 10 seconds before the next affordance, means that build-up time towards this affordance is reduced, thus possibly undesirably surprising the user. Finally, the mix plan is mixed to an audio signal. We employ a window-based approach that produces the audio signal only up to a defined time horizon in the future to avoid wasting computation time on a signal that will be overridden by another update anyways.

3.2.3 Innovative Information Fusion. After playback of a mixed audio signal has started, affordance ETAs continue to be updated at the GPS sample rate, typically revealing discrepancies between the novel and the previous ETA. Generally, as geographic distance to the affordance shrinks, the risk of error in the novel ETA shrinks and its trustworthiness increases. In order to incorporate innovative affordance ETA predictions, we developed a recursive filtering algorithm that continuously fuses the most recent affordance ETA predictions with previous predictions, and determines whether a resynchronization needs to be triggered.

This resynchronization decision is based on the 1) current timestamp, 2) the previously planned ETA of the next affordance, and 3) its novel ETA. We compute a set of descriptive metrics to estimate the current state of the world – that is the environment and the vehicle within that environment – and derive an audio updating strategy.

Figure 6 shows the overall approach. Algorithm 1 gives the in-depth procedure of assessing the state of the system given new ETAs. We *temporize*, i. e., we do not update to new affordance ETAs if there is a deviation between planned and innovative ETA, however it is so far in the future that any update made now is likely to be updated again anyways. This avoids unnecessary but experience-degrading updates. We *neglect* any deviation between planned and

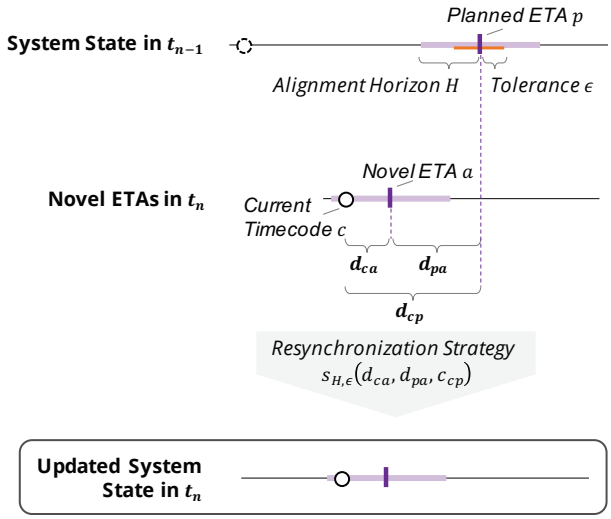


Figure 6: Time-to-arrival estimations are continuously updated by the localization module. To avoid continuous experience-degrading resynchronization updates to the audio signal, SoundsRide must carefully trade off updates against temporal misalignments between the events in music and in the environment. We implement this trade-off in a recursive filtering procedure that continuously fuses novel ETA predictions with former predictions to decide whether a resynchronization needs to be performed.

innovative ETA if it falls below a specified tolerance threshold. We *delay* the next mix event by updating the state of the affordance ETA predictions if the innovative affordance ETA is farther away than the planned affordance ETA. Analogously, we *accelerate* if the innovation suggests an earlier than expected event time. We

Algorithm 1: SoundsRide's ETA fusion algorithm for resynchronization

Input: current timestamp c, novel ETA timestamp a, previously planned ETA timestamp p, alignment horizon H, misalignment threshold ϵ
Output: Updating strategy s for the current timestep

```
s ← undefined
d_cp ← p − c, d_ca ← a − c, d_pa ← a − p
if d_ca ≥ 0 and d_cp ≥ 0 then
    if d_ca > H and d_cp > H then
        // beyond hot zones of novel and planned ETA
        s ← Temporize
    else if d_ca ≤ H and d_cp ≤ H then
        // within hot zone of novel and planned ETA
        if abs(d_pa) < ε then s ← NeglectMisalignment
        else if d_pa ≥ ε then s ← Delay
        else if d_pa ≤ −ε then s ← Accelerate
    else if d_ca ≥ H and d_cp ≤ H then
        // within hot zone of planned ETA
        if abs(d_pa) ≤ ε then s ← NeglectMisalignment
        else s ← Delay
    else if d_ca ≤ H and d_cp ≥ H then
        // within hot zone of novel ETA
        if abs(d_pa) ≤ ε then s ← NeglectMisalignment
        else s ← Accelerate
else
    if d_pa > ε then s ← RedispatchMissedAffordance
    else s ← EndureMissedAffordance
return s
```

endure a misalignment if the affordance is still ahead according to the innovation but should have been passed according to the plan. Only if a threshold is exceeded, we *redispatch* the song snippet that features the musical event. In summary, the updating behavior is controlled by the configurable parameters *misalignment tolerance* and the *alignment horizon*.

4 IMPLEMENTATION

We built SoundsRide in a client-server setup. The *SoundsRide server* is responsible for managing the mixing plan, fusing new affordance ETA information with the mixing plan, computing the mixed audio signal, and forwarding the signal to the vehicle's audio output system via Bluetooth or latency-free AUX. It is implemented in Python. The *SoundsRide client* is responsible for localization using GPS and affordance ETA estimation. It is implemented in Swift for iOS. Interprocess communication is realized through remote procedure calls over HTTP/2 and WiFi using gRPC[5]. Mixing, playback from an audio buffer, real-time visualization, and server communication are all running in different threads on the server. Source code for both the client and the server software is available on GitHub. We chose this prototypical setup to enable rapid implementation on a full-fledged Python server while being able to either consume the GPS information from a common smartphone or from an experimentally tapped vehicle navigation system. For a production deployment, we could imagine both a purely Python-based version running on a vehicle-integrated and Linux-based computation unit, or running as a purely mobile-device-based app, or even running in such a distributed fashion, however delivering the final audio signal over cellular network.

[5]https://grpc.io/

5 EXPLORATORY USER EVALUATION

Participants and Apparatus. To gain insights into how users perceive SoundsRide and which potential for immersion it offers, we conducted a user evaluation ($n = 8, n_{female} = 1, \mu_{age} = 32.9, \sigma_{age} = 8.0, \mu_{km\_pa\_driven} = 18000, \sigma_{km\_pa\_driven} = 11263$). We ran SoundsRide on a common smartphone (in our setup, an Apple iPhone 11 Pro) for GPS localization at 1 Hz and a common laptop (in our setup, a MacBook Pro A2141) for mixing. To measure user reactions to our system, we also tapped gas pedal position signals sent through the vehicle's communication bus system. We recorded the environment with a front-facing camera as well as the sound within the car with a microphone. Figure 7 shows the southwest direction of the test route and its sound affordances. We invited the participants one by one over two days to a location close to the route's start point, starting at approx. 9 am and ending at approx. 21 pm, thus covering different daytime conditions. In one half of the cases, we started with the southwest direction, in the other half, we started the same route in the northeast direction.

Rationale. We choose a field setup over a simulator for two reasons. First, we want to examine user impressions of SoundsRide's real-time capabilities rather than the idea under ideal conditions. Second, as SoundsRide does not assist in a driving task but aims at evoking subjective impressions, we give weight to the fidelity of the evaluation and want to ensure full visual, physical and auditory contrasts, e. g., at the tunnel entrance affordance. On the other hand, a simulation might entail questions on the real-world transferability.

Following Qin et al. [28] for immersion in digital gaming, control, concentration and comprehension serve as explanatory factors for immersion. In particular, comprehension is a necessary condition for immersion. Therefore, first, we wanted to understand how much sense SoundsRide makes to users intuitively, i. e., without exactly knowing its functionality. Once we have investigated this necessary condition, we evaluate which potential for immersion it offers to users.

Therefore, after a baseline part, the subsequent procedure was composed of two main segments, namely 1) investigating perceptibility, and 2) investigating the potential for immersion. The rationale for conducting separate studies to investigate perceptibility and potential for immersion is twofold:

(1) By not previously telling participants what will happen, we do not expose them to confirmation bias when investigating how comprehensible the system is to users and what impression it evokes in them
(2) By separating both study aspects, we can more reliably isolate insights concerning technical design aspects from insights concerning conceptual design aspects. If users were to undertake only an immersion study, in the case of a participant showing no indications of immersion, it is unclear whether this is *because of not* perceiving the system or *despite* perceiving it.

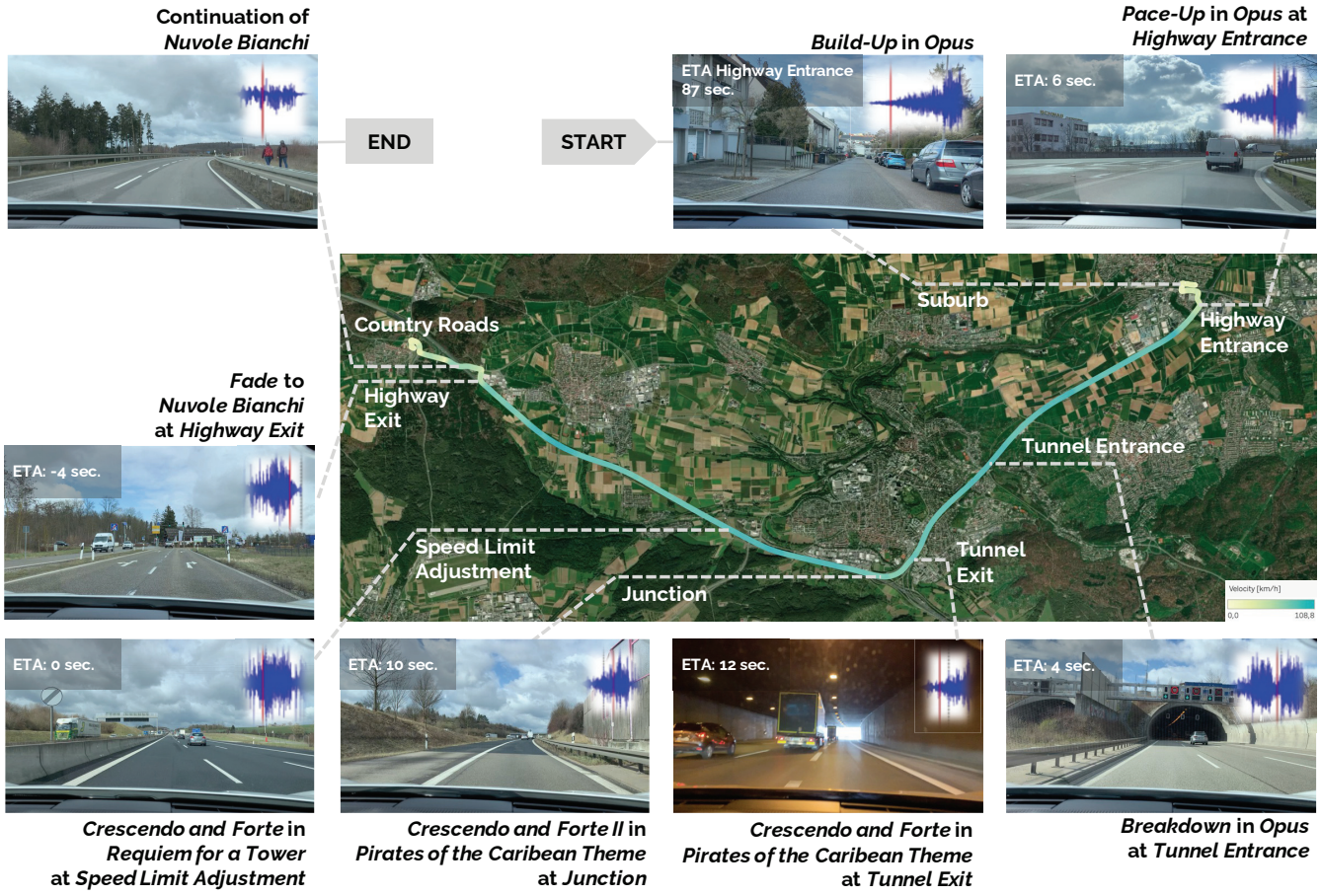


Figure 7: Route and Affordances for our User Evaluation in Southwest Direction. The red marker in the audio signal indicates the current position in the audio track. We selected a test route of 15.9 km length that takes about 12 to 15 minutes on average at average traffic with 4 affordances in northeast direction and 6 affordances in southwest direction shown here. We chose an electric vehicle from our institution. We recorded reference GPS trajectories in both directions. We employed SoundsRide's automatic mixing mode during the studies, but for comparability reasons, fixed the sequence of songs. Misalignment tolerance was set to 1000ms and the alignment horizon to 15s.

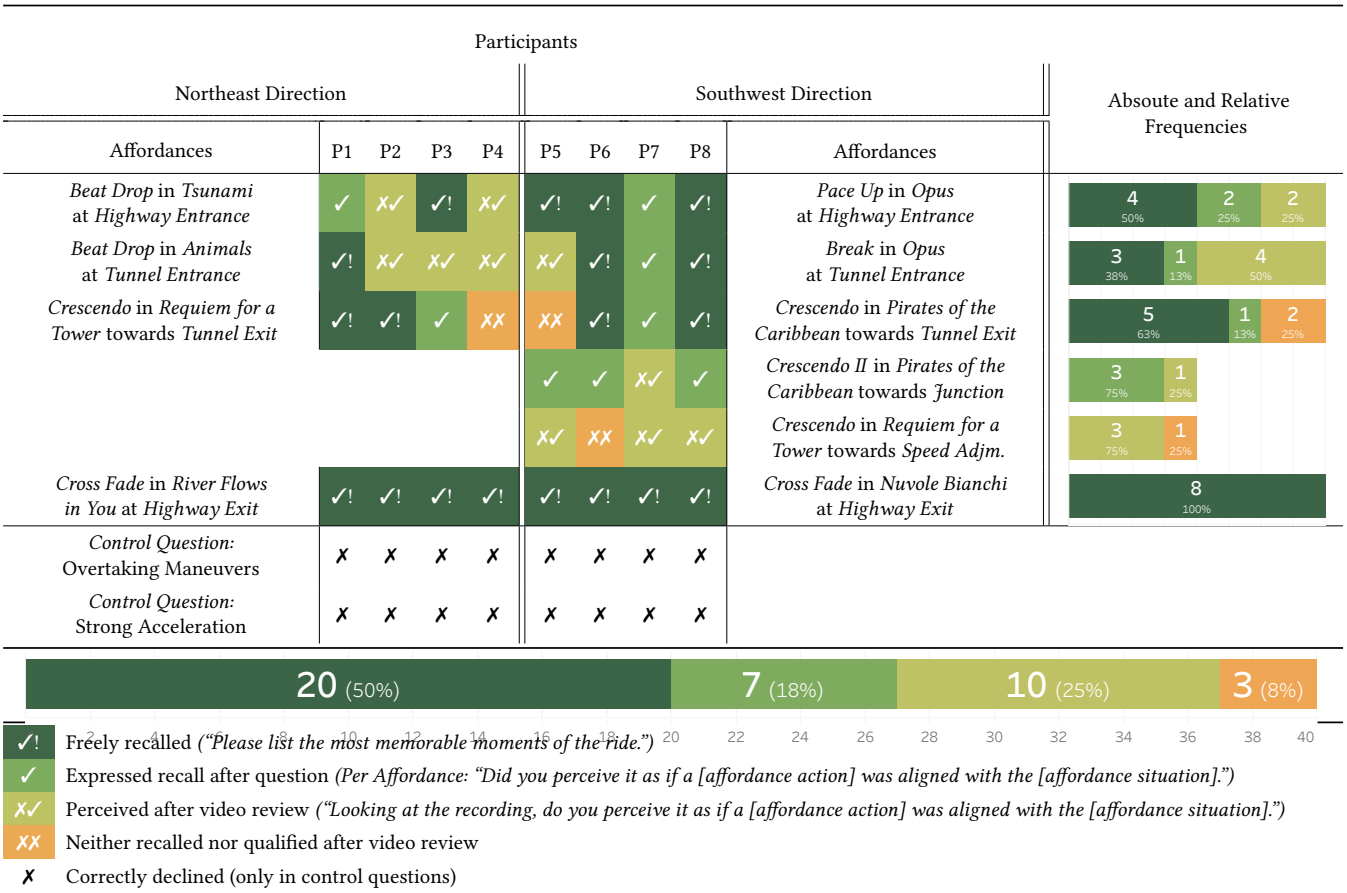
5.1 Study Segment 1: Investigating Perceptibility

5.1.1 Procedure. To investigate the system's perceptibility, each participant was shortly briefed with respect to the abstract capability of the system – that is that the system is "able to align certain events in the music with certain events during the ride". They were *not* informed about the concrete types of sound affordances supported. Further, we asked participants to "think out loud" and state their assumptions about what they believe the system is doing. We navigated participants by verbal instructions. After arriving at the end of the test route, we ask a cascade of questions to understand the system's percepbility to users, asking for free recall of affordances, then cued recall to video recall or no qualification as a music-environment-aligned affordance. The caption in Table 1 elaborates on the details of the procedure.

5.1.2 Results and Discussion. Table 1 shows the data collected during the study and aggregations thereof. 67.5 % of all affordances were either recalled freely or recalled after a cue question. 25.0 % of all affordances were not recognized by participants during the ride, however were qualified as such when reviewing the ride on the recorded video.

Both P1 and P4 did not remember a music synchronization at the highway entrance, but expressed puzzlement when looking at the recording and asked themselves why they did not notice the well-aligned and lucid beat drop in Tsunami. Similarly, P8 did not remember music synchronization with the speed limit adjustment, however suddenly remembered that they were very focused on overtaking a "white Tesla". In the later study segment on immersion, P8 also stated that they felt the high-energy music during the highway entrance was not supporting their mindfulness. Possibly, P1 and P4 blocked out the music to direct their attention to merge

Table 1: After arriving at the end of the test route, we first asked the participant to list the most memorable events from the ride. Then, for each non-mentioned affordance as well as for two control items (i. e., environment events SoundsRide does not take into account, namely overtaking maneuvers and strong acceleration), we asked participants directly whether they perceived it (e. g., "From what you remember, did you perceive it as if a beat drop was aligned with the tunnel entrance?"). If answered negatively, we showed the participant the front-facing video recording of the ride including sound, and asked again shortly after the sound affordance took place (e. g., "Looking at the video recording, do you perceive it as if a beat drop was aligned with the tunnel entrance?"), to understand whether the reason for not remembering was not perceiving it *just in that moment* while during, e. g., due to high cognitive load, or whether the reason lied in generally not appreciating the affordance, e.g., because of a too large time delta between affordance trigger and affordance action or too much subjective indifference towards the contrast in music or environment.



on the highway. *From this we conclude* that perceptibility must not be an overarching objective of the system. On the contrary, the system should fade into the background during critical segments such as highway entrances and only become noticeable again once the segment is finished. This can be realized by either moving the affordance location forward, selecting the music accordingly or both.

While P3, P4, and P7, missed two affordances each, P1 and P6 did not miss any. This does not correlate to the data on driving experience in age or km driven per year. However, as 6 out of 10 misses happened at the NE tunnel entrance and SW speed adjustment, both not visible from far as the former is hidden behind the curve and the latter is only a small road sign, we hypothesize that users miss affordances without sufficient time for anticipating or following the build-up in music. As inter-song events do not benefit from anticipation and are more distinct, they might be better suited for such locations.

In 7.5 % of affordances, participants rejected the notion that the music was adapting synchronously to the environment also after looking at the recording. In the case of P5 approaching the tunnel exit, a loss of the GPS signal inside the tunnel did result in a misalignment between music and tunnel exit of approx. 5 seconds, thus failing to keep music and environment in sync. In the case of P4 approaching the tunnel exit, the participant stated that the change in music at the tunnel exit was not enough to make a difference. Similarly, in the case of P6 approaching the speed

limit adjustment, the same audio track was rejected for being too monotonous. However, the tunnel exit with the same orchestral piece was remembered by other participants 3 times. *From this we conclude* that the subjective feeling of energy a build-up or climax evokes is a significant determinant in the experience of SoundsRide.

All participants could freely recall the fade from orchestral music to piano music towards the highway exit. The speed limit adjustment was not recalled freely once. Surprisingly, *all* of the control item questions were correctly rejected by participants, strengthening our confidence in the validity of the answers overall.

Overall, we conclude that SoundsRide is well comprehensible to users but subjectivity of musical perception and mental load reduce perceptibility, first, indicating the need for customization of the song selection per user, and second, underlining the necessity to analyze the implications of SoundsRide for driving safety.

5.2 Study Segment 2: Investigating the Potential for Immersion

In the subsequent *Potential for Immersion study*, we wanted to understand whether the system allows users to immerse deeper in the music or in the environment than without SoundsRide. More specifically, we want to understand the experience of 1) synchronicity, 2) affordances, 3) mixing, 3) the effect on driving safety and 5) overall immersion.

5.2.1 Procedure. We disclosed the affordances (using the word "event") taken into account by SoundsRide to align the music for the participants and asked them to drive in the opposite direction of the route for the perceptibility study. We announced the next affordance 30 to 60 seconds before it took place. Once again, we asked participants to "think out loud" and describe what they are thinking or feeling. Again, after arriving at the end of the test route, we employed a questionnaire.

5.2.2 Results and Discussion.

Experience of Synchronicity. Regarding *micro-synchronicity*, we found that participants judged temporal misalignments subjectively. While P4 negatively commented on the forte at the speed limit adjustment being off by approx. 1.5 seconds, P2 commented positively at the same affordance and offset. 4 out of 8 participants stated that they were very much (1x) or much (3x) disappointed when the system would miss an expected affordance by more than a second. Two participants stated the question was not applicable given they could not remember such situations. Two participants expressed little to no disappointment, indicating little emotional attachment. Also, we found that affordance positioning is user-dependent in some parts. While all participants gave positive feedback concerning the fade to the piano piece at the highway exit, P5 added that it was taking place too early in their mind as they drive the curve diverging from the highway at a higher speed than is adequate for the calming effect of the piano. Regarding *macro-synchronicity*, P4, P7, and P8 noted that the continuation of the orchestral "Requiem for a Tower" lost its fit to the environment continuously after the respective event (speed limit adjustment or tunnel exit respectively) had passed. In particular, P8 said that the continuation of the song would naturally lead to music events that are decoupled from the environment, even though they would not perceive this as detrimental to the experience. P1, P2, P3, P4, and P7 noted that overall synchronicity was impaired for a short duration at a traffic light where the music was already building up to a beat drop while still waiting at the red traffic light.

Experience of Affordances. Situation-wise, P5 and P6 dismissed the notion that a tunnel is an event during the ride worth synchronizing the music for. P6 reported that only the highway exit with the synchronized fade to piano music captured their attention while none of the other songs or pieces "did anything to me". However, all other participants deemed the affordance situations fitting and interesting. *Action-wise*, all participants favored the cross-fade to a piano piece at the highway exit. However, P6 stated, except for the piano piece, no piece or song was in the domain of their music preference and thus interesting to them. As a result, while they would auditorily recognize environment-triggered developments in the music, they did have the interest to actively follow it. P5 stated that driving is a "mechanical task that needs to be done" and not more, and hence "they block out the environment as far as it is not needed for driving safely". Therefore, the desire to listen to a certain song or is not a function of the ride or the location, but of their current mood. Except for P5 and P6, participants also liked the beat-drop in Animals (northeast direction) and the break in Opus (southwest direction) towards the tunnel entrance, however the beat-drop evoked more vivid feedback. Participants also expressed liking towards the orchestral crescendo at the tunnel exit. The continuous progression in tempo in Opus and the pace-up at the highway entrance was generally well-appreciated, however – as described in the results on study segment 1 – P8 expressed that they would have preferred music that enables them to better concentrate on the merge.

Experience of Mixing. 7 out of 8 participants stated that the transitions from one song to the next, and the concomitant less-than-usual length of the song did not invoke stress, while one participant stated to experience light stress. However, 5 out of 8 participants reported that adaptions of an already playing song were annoying. These were the participants, who also noted that the traffic light was a source of asynchronicity, namely between a build-up in music and an unvaried environment. In particular, P2 described it "as being in a loop where the music gets faster, then gets slower, then gets faster and slower again while nothing actually happens". The described system behavior is a result of the system resetting the track to avoid a premature event in the music. However, this phenomenon generally did not lead to confusion, considering only 1 out of 8 participants agreed they felt confused by the system. Participants did not comment or call out on audio track modifications that were taking place very shortly, approx. 2 seconds, before the affordance situation took place. From this, we conclude that improving affordance ETAs on a larger scale are of first priority, e. g., by taking traffic lights into account during planning, while avoiding modifications in the last one or two seconds are of subordinate priority.

Effect on Driving Safety. It is of crucial importance that SoundsRide does not impair driving safety through mechanisms of adverse incentives or distraction. 2 out of 8 participants responded with

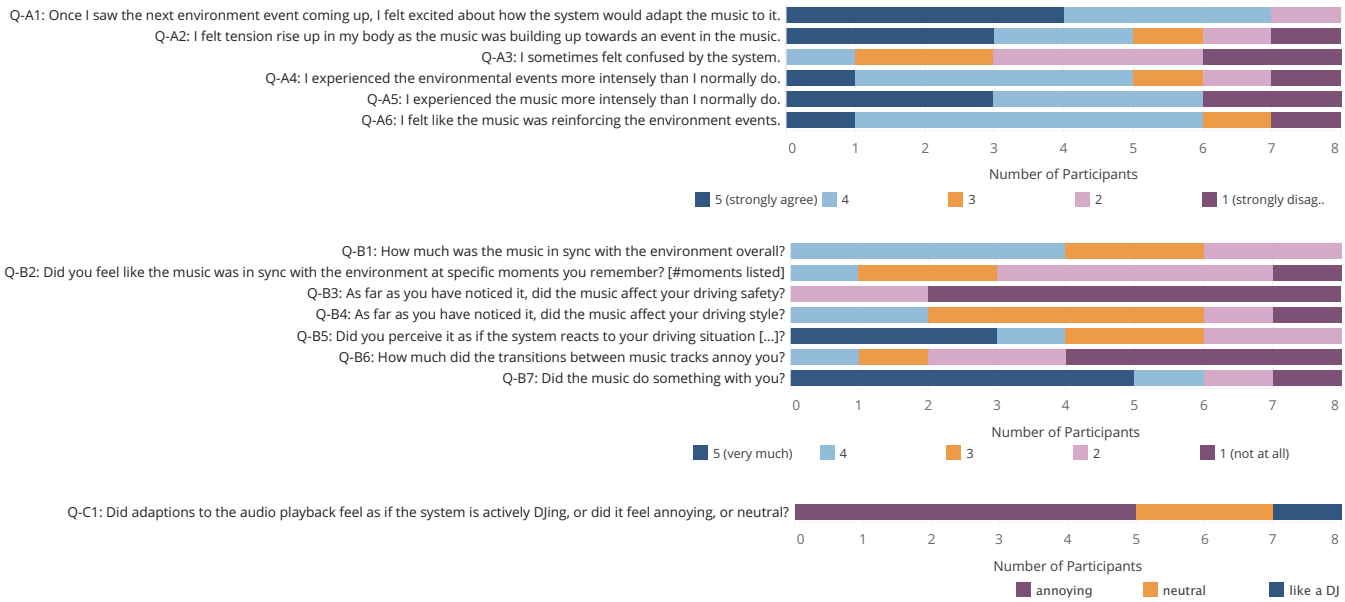


Figure 8: Participant's responses to our immersion questionnaire. After arriving at the end of the test route in study segment 2 on immersion, we employed a questionnaire to understand users' experiences in terms of affordances, mixing, and synchronicity, to understand the subjectively perceived effect on driving safety, and to understand the overall potential for immersion.

a value of 2 on a scale from 1 (not at all) to 5 (very much) to the question "As far as you have noticed, how much did the music affect your driving safety". One of these two participants has indicated the same in the questionnaire precursory to the ride to the same question, however in the general context of in-car music. This means SoundsRide did not negatively impact the participant compared to a baseline, but also did not manage to eliminate this by guiding focus towards the environment. Further, P8 said that the pace-up on the highway entrance was opposite to their need for deep focus. 6 out of 8 participants responded with a value of 1. P7 stated that they would normally drive too fast after driving off the highway, however due to the switch to piano music, they'd prefer "just coasting along". P2 stated that the music helped them concentrate by directing their focus on the environment and prepare for the next driving maneuver. From these answers, we conceive that SoundsRide can *negatively influence* safety by incentivizing users to increase velocity with increasing energy in music, e. g., at the highway entrance, or by distracting them from the driving situation. On the other hand, we conceive that SoundsRide can *positively influence* safety by incentivizing users to decrease velocity with decreasing energy in music, e. g., slowing down according to calm piano music at the highway exit, or by increasing environment awareness, e. g., so that users slow down early enough before the highway exit as the song starts fading out. *Overall*, we derive the hypothesis that SoundsRide has both the potential to slightly increase as well as slightly decrease driving safety. As a consequence, further research needs to investigate which properties in the situation and user prejudice or contribute to safety and how to eliminate or foster these effects, e.g., through careful positioning of affordances before or behind precarious points on the route.

Overall Immersion. Especially building on Brown and Cairns [6] as well as Qin et al. [28], Georgiou and Kyze [14] propose constructs to be measured for immersion as well a questionnaire to operationalize these constructs in a study. Three consecutive levels of immersion are distinguished: engagement, engrossment, and total immersion. Classically, Level 1, engagement, represents the basic level and is characterized by the will to interact with a system. Level 2, engrossment, is characterized by an emotional attachment. Level 3, total immersion, is characterized by presence, which of course cannot be a goal for driver-directed applications. While Georgiou and Kyze's questionnaire is designed for evaluating see-through AR games, we took it as as a starting point for our in-car audio AR evaluation questionnaire. *Considering the aggregated data*, 7 out of 8 participants agreed, some strongly, that they were excited about how the music would adapt to the next environment event (i. e., the affordance situation), once they saw it coming up. 5 out of 8 agreed, some strongly, that they felt the tension of a build-up in music towards an event physically. 6 out of 8 agreed, one strongly, that the music was reinforcing the environment events. 6 out of 8 agreed, some strongly, that they experienced the music more intensely than they normally do. 5 out of 8 participants agreed, some strongly, that they experienced the affordance situations more intensely than normally. Of course, attention bias due to the study setup mandates to interpret responses to both former questions cautiously. *Considering the individual data*, as described above, P6 stated multiple times that the system did not at all capture their

attention or emotion because the music did not mean anything to them. However, the participant was open to retrying the system should it support custom music selections. On the contrary, P5 rejected the premise of environment-induced music overall. On the other hand, P3 stated they "felt like being in a car advertisement". P2 stated that they "didn't know how spectacular merging on a highway could feel". P1 stated that the system put them "in a spirit of optimism and anticipation for the highway". On a more general note, the participants' statements reveal that the quality of the experience is subject to a range of context variables. Concerning *traffic*, P3 stated that the energetic build-up in music at the highway entrance evoked the "urge to accelerate", but the "traffic wouldn't allow it". Concerning *aesthics of the scene*, P7 stated that they felt as if they were "experiencing something very special, but the grey trucks bring me down to earth again". Concerning *weather*, P7 expressed the wish that "the rain should reflect in music to convey a sense of melancholy". Concerning *daytime*, P4 said that the "gloomy music [of Requiem for a Tower] fits the mood". Concerning *vehicle type*, P3 indicated that the "electric engine works well with the peaceful piano at slow down" behind the highway exit, whereas P6 noted that "a combustion engine fits the emotional experience better".

Overall, from the aggregate statistics and the individual responses, we conclude that SoundsRide does offer the potential for immersive experiences through emotional attachment, assuming a user is not rejecting the system's fundamental premise of affordance-based music. However, we note that profile-based or personalized music could also capture the attention of users whose music preferences were not covered in the selection, and that SoundsRide's experience is dependent on a variety of context variables which the system could potentially account for in the future.

6 TECHNICAL EVALUATION

Since SoundsRide focuses on affordance synchronization, the main metric to evaluate its overall technical performance is given through temporal misalignment, i. e., the temporal distance between the mix event and the environment event. However, as temporal misalignment can be traded-off against re-synchronization updates to the audio signal, we also consider the number of audio signal updates before (speed-up or delay update) and after the affordance situation (redispatch).

6.1 Procedure

We recorded the rides during the exploratory user evaluation with a front-facing camera and incl. sound. Due to the nature of the setup, these rides feature different users, driving styles, and environment conditions, in particular rush hours in the morning and the afternoon, and free roads in the evening. We annotated the videos using a video editing software for all 48 affordances with the temporal misalignment and the number of audio track per affordance.

6.2 Results

Figure 9 shows the distribution of the temporal misalignment ($\mu = 942\,ms$, $\sigma = 1813\,ms$) and the number of update per affordance across all 8 rides.

Except for the tunnel exit affordance, all events were in a range of no more than four seconds. However, in approx. half of the cases (47.7 %, including the tunnel exit affordance) SoundsRide is able to ensure synchronicity within a time horizon of $\pm 1.1\,s$ and with no more than 1 track update in the 15 seconds before or after the affordance location is passed. In 77.7 % of cases, SoundsRide is able to ensure synchronicity with a maximum misalignment of $\pm 1.9s$ and a maximum number of 2 updates to the audio track. All rides were parameterized with a misalignment tolerance of $1000ms$ and an alignment horizon of $15s$.

The striking elongated distribution at the tunnel exit results from the decreased accuracy of the GPS signal in the tunnel. The number of updates is higher at average at the highway entrance and at the junction than at the tunnel entrance, as the exact timing of these affordance situations is more influenced by traffic and thus more difficult to predict from the reference trajectory. At the highway exit, we deliberately employed a long-running cross-fade of 10 seconds aligned with the exit lane branching from the highway without a point-exact location. In all rides, the affordance action was taken on the exit lane, thus meeting this condition.

7 LIMITATIONS AND FUTURE WORK

Mixing. The most frequent criticism of participants in the evaluation concerned the track resetting procedure. While we employ a user-parametrizable filtering procedure to minimize the number of noticeable mix modifications given new ETA information, updates that do take place are typically noticeable to users. By analyzing a song, e. g., detecting its BPM and segmenting it in individual bars and phases, music-theoretically more sound updating procedures to achieve temporal alignment can be imagined, e. g., bar or phase repetitions. In particular, employing BPM stretching in an acoustically safe range while simultaneously correcting for pitch could induce a very interesting, possibly unnoticeable, audio effect. That is, pushing the gas pedal or the brake pedal will indirectly lead to speeding up or slowing down the music in an effort of the system to reach the aligned event in the music earlier or later than originally planned.

Synchronicity. Macro-synchronicity can be increased by manually annotating traffic lights or inferring them from rests in the GPS trajectory during route recording. Then, the system can ensure to only transition to the next song featuring the next musical event, once the traffic light has been passed, thus further reducing track resets from unexpected delays or speed-ups. To improve micro-synchronicity in general, reference-based ETA prediction could be enriched with speed awareness and acceleration awareness. To improve micro-synchronicity at tunnel exits specifically, by integrating SoundsRide with a vehicular information system, the GPS localization from a smartphone could be fused with odometric information from the vehicle's wheel speed sensor, compensating the loss of the GPS signal in the tunnel.

Affordance Situations. We have designed SoundsRide for full flexibility on location-bound affordance situations. By first recording a reference route, users, community members, or third-party providers can freely annotate location-bound affordances. At the same time, the reference route is taken as a basis for predicting

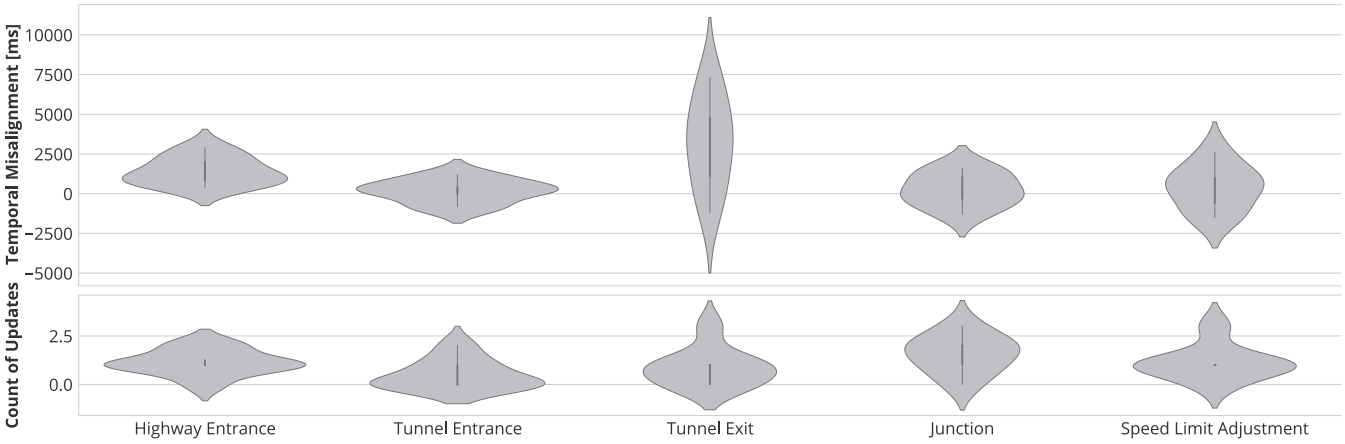


Figure 9: Distribution of temporal misalignment and number of updates per affordance across 8 rides, each with a misalignment tolerance of 1000 ms and an alignment horizon of 15 s. In 77.7 % of cases (73.7 % without the highway exit affordance), SoundsRide is able to ensure synchronicity with a maximum misalignment of ±1.9s and a maximum number of 2 updates to the audio track.

affordance ETAs. However, extracting location-bound affordances such as highways, tunnels, hill crests, etc. from and predicting their ETAs with mapping services such as OpenStreetMap could enable SoundsRide for one-step usage without manual annotation efforts just by entering the route to be driven. Going even further, continuous automatic inference of the most probable next route segment the user will drive, even without explicit entry of a destination, might allow determining the next affordance on-the-fly.

Affordance Actions. SoundsRide works best with songs featuring high contrastiveness and mixability. Therefore, EDM, orchestral, jazz-only, and piano-only music works particularly well. Nu metal, pop, rock, and pop-rock also work well in terms of contrasts, however, mixing runs the risk of pulling apart vocals when skipping or rewinding for resynchronization and fading. To make SoundsRide more general, future work can explore three strategies. First, by expanding the current structural awareness to vocal awareness, i. e., knowledge of timecodes of verses, the system could jump to dedicated markers, thereby keeping vocals together. Second, by further investigating how to improve micro-temporal estimation, the number of resynchronization updates and thus the risk of cutting vocals is reduced. Third, the adoption of BPM stretching as presented above could avoid jumps due to resynchronization entirely. In order to feature songs of different genres in the same mix, research into the automation of music-theoretically informed harmonic song transitions [35] gains in importance. While it is already perfectly possible to create a custom song database, this requires annotating each song with intra-song events and defining the mapping between sound affordance situations and the respective intra-song event. However, adopting approaches of structural segmentation [25] to detect segment boundaries and classes in songs could allow users to simply hand-over a playlist to the system before the next ride, thus conveniently enabling personalization and likely increasing the potential for immersion. Additionally, including further affordance actions such as balancing the music from front to back speakers when entering a tunnel, or panning the music from left to right when entering a highway based on spatial audio, or even employing other modalities such as actuated seats, could enrich SoundsRide's experience and increase its potential for immersion.

Adaption to Automated Driving. The trend towards driving automation affects a variety of the system's aspects. For example, driving scenario detection could enable dynamic affordances that react on-the-fly to triggers in traffic, e. g., allowing to inject energetic music into the mix when overtaking or being overtaken. Visualizing affordance situations in the driving scene display could enable possibly interactive monitoring. Advanced depth-sensing might allow correcting ETAs on short term based on the environment scans, e. g., when detecting a forest boundary. Overall, ETA accuracy might benefit from automated driving due to improved predictability in the driving behavior. On the other hand, automation might impair the UX as the driver possibly starts shutting out the environment. Therefore, future work might also explore novel interaction patterns, e. g., where a user co-creates the mix plan by scheduling affordances on short-term notice and on-the-fly based on music choices offered by the system. Also, we imagine the system might enable audio-based in-car games, e. g. passengers guessing the current scene blindfolded based on music only.

Domain Adaption. Finally, we see potential to transfer SoundsRide to other means of locomotion in general and bicycles in particular, e. g., in order to synchronize energetic music with uphill segments on the ride, or synchronize the transition from bicycle trails to public roads.

8 CONCLUSION

We presented SoundsRide, an in-car audio augmented reality system that synchronizes high-contrast events in music with high-contrast events in the environment. Our core technical contribution lies in predictive real-time music mixing, enabled by a novel approach

comprising affordance ETA prediction, mix planning, and innovative information fusion.

After positioning SoundsRide in the design space of affordance-based music, we describe our technical approach and its implementation. Given the estimated temporal distance to location-bound affordances along the ride, we heuristically schedule a cost-aware music mix with intra-song and inter-song events that are temporally aligned with the affordances. Then, by continuously piping updated temporal affordances distances through a recursive filter, we determine any necessary updates to the audio signal while trading off manipulations of the audio signal against misalignments between music and environment.

On the one hand, using SoundsRide's automatic mixing mode, users are enabled to align events in any set of annotated songs with configured affordance situations along their ride, thus making it applicable for everyday rides. On the other hand, using a predefined mix possibly offered by a regional provider such as a national park service or a local tourism association, we also envision SoundsRide to further increase immersion for particularly captivating or scenic routes.

In our quantitative evaluation of SoundsRide's performance, we find that SoundsRide can convincingly ensure synchronicity of affordance situations in the environment and affordance actions in the music. In our qualitative evaluation, we find that SoundsRide's affordances as well as its mixing and synchronicity properties are well-received, thus enabling suspenseful and engrossing experiences, and that driving safety can be either slightly increased or decreased, depending on the mix and user.

REFERENCES

[1] Robert Albrecht, Riitta Väänänen, and Tapio Lokki. 2016. Guided by music: pedestrian and cyclist navigation with route and beacon guidance. *Personal and Ubiquitous Computing* 20, 1 (2016), 121–145. https://doi.org/10.1007/s00779-016-0906-z

[2] Linas Baltrunas, Marius Kaminskas, Bernd Ludwig, Omar Moling, Francesco Ricci, Aykan Aydin, Karl Heinz Lüke, and Roland Schwaiger. 2011. InCarMusic: Context-aware music recommendations in a car. *Lecture Notes in Business Information Processing* 85 LNBIP (2011), 89–100. https://doi.org/10.1007/978-3-642-23014-1_8

[3] Benjamin B. Bederson. 1995. Audio augmented reality. (1995), 210–211. https://doi.org/10.1145/223355.223526

[4] Matthias Braunhofer, Marius Kaminskas, and Francesco Ricci. 2013. Location-aware music recommendation. *International Journal of Multimedia Information Retrieval* 2, 1 (2013), 31–44. https://doi.org/10.1007/s13735-012-0032-2

[5] Warren Brodsky. 2001. The effects of music tempo on simulated driving performance and vehicular control. *Transportation Research Part F: Traffic Psychology and Behaviour* 4, 4 (2001), 219–241. https://doi.org/10.1016/S1369-8478(01)00025-0

[6] E Brown and P Cairns. 2004. A grounded investigation of immersion in games. *Proc. CHI EA '04* (2004), 31–32. http://discovery.ucl.ac.uk/55390/

[7] Gary Burnett, Adrian Hazzard, Elizabeth Crundall, and David Crundall. 2017. Altering speed perception through the subliminal adaptation of music within a vehicle. *AutomotiveUI 2017 - 9th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, Proceedings* (2017), 164–172. https://doi.org/10.1145/3122986.3122990

[8] Zhiyong Cheng and Jialie Shen. 2016. On effective location-aware music recommendation. *ACM Transactions on Information Systems* 34, 2 (2016). https://doi.org/10.1145/2846092

[9] Karen Collins. 2009. An introduction to procedural music in video games. *Contemporary Music Review* 28, 1 (2009), 5–15. https://doi.org/10.1080/07494460802663983

[10] Stefan K. Ehrlich, Kat R. Agres, Cuntai Guan, and Gordon Cheng. 2019. A closed-loop, music-based brain-computer interface for emotion mediation. *PLoS ONE* 14, 3 (2019), 1–24. https://doi.org/10.1371/journal.pone.0213516

[11] Greg T. Elliott and Bill Tomlinson. 2006. PersonalSoundtrack: Context-aware playlists that adapt to user pace. *CHI EA '06: CHI '06 Extended Abstracts on Human Factors in Computing Systems* (2006), 736–741. https://doi.org/10.1145/1125451.1125599

[12] Seyedeh Maryam Fakhrhosseini, Steven Landry, Yin Yin Tan, Saru Bhattarai, and Myounghoon Jeon. 2014. If you're angry, turn the music on: Music can mitigate anger effects on driving performance. *AutomotiveUI 2014 - 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, in Cooperation with ACM SIGCHI - Proceedings* (2014). https://doi.org/10.1145/2667317.2667410

[13] Emma Frid, Celso Gomes, and Zeyu Jin. 2020. Music Creation by Example. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–13. https://doi.org/10.1145/3313831.3376514

[14] Yiannis Georgiou and Eleni A. Kyza. 2017. The development and validation of the ARI questionnaire: An instrument for measuring immersion in location-based augmented reality settings. *International Journal of Human Computer Studies* 98, December 2015 (2017), 24–37. https://doi.org/10.1016/j.ijhcs.2016.09.014

[15] Gerhard Johann Hagerer, Michael Lux, Stefan Ehrlich, and Gordon Cheng. 2015. Augmenting affect from speech with generative music. *CHI EA '15: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* 18 (2015), 977–982. https://doi.org/10.1145/2702613.2732792

[16] Patrick Helmholz, Sebastian Vetter, and Susanne Robra-Bissantz. 2014. AmbiTune: Bringing context-awareness to music playlists while driving. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8463 LNCS (2014), 393–397. https://doi.org/10.1007/978-3-319-06701-8_32

[17] Patrick Helmholz, Edgar Ziesmann, and Susanne Robra-Bissantz. 2013. Context-awareness in the car: Prediction, evaluation and usage of route trajectories. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7939 LNCS (2013), 413–419. https://doi.org/10.1007/978-3-642-38827-9_30

[18] Xiping Hu, Junqi Deng, Jidi Zhao, Wenyan Hu, Edith C.H. Ngai, Renfei Wang, Johnny Shen, Min Liang, Xitong Li, Victor C.M. Leung, and Yu Kwong Kwok. 2015. SAfeDJ: A crowd-cloud codesign approach to situation-aware music delivery for drivers. *ACM Transactions on Multimedia Computing, Communications and Applications* 12, 1 (2015). https://doi.org/10.1145/2808201

[19] Haikun Huang, Michael Solah, Dingzeyu Li, and Lap Fai Yu. 2019. Audible panorama: Automatic spatial audio generation for panorama imagery. *CHI '19: Proceedings of CHI Conference on Human Factors in Computing Systems* Chi (2019), 1–11. https://doi.org/10.1145/3290605.3300851

[20] Victor Kaptelinin and Bonnie Nardi. 2012. Affordances in HCI: Toward a mediated action perspective. *CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), 967–976. https://doi.org/10.1145/2207676.2208541

[21] Michael Krzyzaniak, David Frohlich, and Philip J.B. Jackson. 2019. Six types of audio that DEFY reality!: A taxonomy of audio augmented reality with examples. *AM'19: Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound* (2019), 160–167. https://doi.org/10.1145/3356590.3356615

[22] Michael Z. Land and McConnell Peter N. 1994. Method and apparatus for dynamically composing music and sound effects using a computer entertainment system.

[23] Wei Po Lee, Chun Ting Chen, Jhih Yuan Huang, and Jhen Yi Liang. 2017. A smartphone-based activity-aware system for music streaming recommendation. *Knowledge-Based Systems* 131 (2017), 70–82. https://doi.org/10.1016/j.knosys.2017.06.002

[24] Mark McGill, Stephen Brewster, David McGookin, and Graham Wilson. 2020. Acoustic Transparency and the Changing Soundscape of Auditory Mixed Reality. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–16. https://doi.org/10.1145/3313831.3376702

[25] Oriol Nieto and Juan Pablo Bello. 2016. Systematic exploration of computational music structure research. *ISMIR '16: Proceedings of the 17th International Society for Music Information Retrieval Conference* (2016), 547–553.

[26] Adrian C. North and David J. Hargreaves. 1999. Music and driving game performance. *Scandinavian Journal of Psychology* 40, 4 (1999), 285–292. https://doi.org/10.1111/1467-9450.404128

[27] Peter Peerdeman. 2006. Sound and Music in Games. *And Haugehåtveit, O* April (2006), 1–18. http://www.peterpeerdeman.nl/vu/ls/peerdeman_sound_and_music_in_games.pdf

[28] Hua Qin, Pei Luen Patrick Rau, and Gavriel Salvendy. 2009. Measuring player immersion in the computer game narrative. *International Journal of Human-Computer Interaction* 25, 2 (2009), 107–133. https://doi.org/10.1080/10447310802546732

[29] Steve Rubin and Maneesh Agrawala. 2014. Generating emotionally relevant musical scores for audio stories. *UIST '14: Proceedings of the 27th annual ACM symposium on User Interface Software and Technology* (2014), 439–448. https://doi.org/10.1145/2642918.2647406

[30] Steve Rubin, Floraine Berthouzoz, Gautham J. Mysore, Wilmot Li, and Maneesh Agrawala. 2012. UnderScore: Musical underlays for audio stories. *UIST '12: Proceedings of the 25th annual ACM symposium on User interface Software and Technology* (2012), 359–366.

[31] Haruki Sato, Tatsunori Hirai, Tomoyasu Nakano, Masataka Goto, and Shigeo Morishima. 2015. A music video authoring system synchronizing climax of video clips and music via rearrangement of musical bars. *ACM SIGGRAPH 2015 Posters, SIGGRAPH 2015* (2015), 2010. https://doi.org/10.1145/2787626.2792608

[32] Norma Saiph Savage, Maciej Baranski, Norma Elva Chavez, and Tobias Höllerer. 2012. I'm feeling LoCo: A location based context aware recommendation system. *Lecture Notes in Geoinformation and Cartography* 199599 (2012), 37–54. https://doi.org/10.1007/978-3-642-24198-7_3

[33] Eldon Schoop, James Smith, and Bjoern Hartmann. 2018. HindSight: Enhancing spatial awareness by sonifying detected objects in real-time 360-degree video. *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* 2018-April (2018), 1–12. https://doi.org/10.1145/3173574.3173717

[34] Marjolein D. van der Zwaag, Chris Dijksterhuis, Dick de Waard, Ben L.J.M. Mulder, Joyce H.D.M. Westerink, and Karel A. Brookhuis. 2012. The influence of music on mood and performance while driving. *Ergonomics* 55, 1 (2012), 12–22. https://doi.org/10.1080/00140139.2011.638403

[35] Len Vande Veire and Tijl De Bie. 2018. From raw audio to a seamless mix: creating an automated DJ system for Drum and Bass. *Eurasip Journal on Audio, Speech, and Music Processing* 2018, 1 (2018). https://doi.org/10.1186/s13636-018-0134-8

[36] Yujia Wang, Wei Liang, Wanwan Li, Dingzeyu Li, and Lap-Fai Yu. 2020. Scene-Aware Background Music Synthesis. (2020), 1162–1170. https://doi.org/10.1145/3394171.3413894

[37] Huiying Wen, N. N. Sze, Qiang Zeng, and Sangen Hu. 2019. Effect of music listening on physiological condition, mental workload, and driving performance with consideration of driver temperament. *International Journal of Environmental Research and Public Health* 16, 15 (2019). https://doi.org/10.3390/ijerph16152766

[38] Zach Whalen. 2004. Play along - An approach to videogame music. *Game Studies* 4, 1 (2004), 1–28.

[39] Sebastian Zepf, Javier Hernandez, Monique Dittrich, and Alexander Schmitt. 2019. Towards empathetic car interfaces: Emotional triggers while driving. *CHI EA '19: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), 1–6. https://doi.org/10.1145/3290607.3312883

[40] Yueyan Zhu, Ying Wang, Guofa Li, and Xiang Guo. 2016. Recognizing and releasing drivers' negative emotions by using music: Evidence from driver anger. *AutomotiveUI '16 Adjunct: Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2016), 173–178. https://doi.org/10.1145/3004323.3004344