INTRODUCTION OF A COMPUTATIONAL MODELLING APPROACH TO AUDITORY DISPLAY RESEARCH: CASE STUDIES USING THE QN-MHP FRAMEWORK

Myounghoon Jeon, Chihab Nadri

Mind Music Machine Lab,
Virginia Tech,
Blacksburg, USA
{myounghoonjeon, cnadri}@vt.edu

Yiqi Zhang

Human-Technology Interaction Lab,
Pennsylvania State University,
University Park, USA
yuz450@psu.edu

ABSTRACT

For more than two decades, a myriad of design and research methods have been proposed in the ICAD community. Neurological methods have been presented since the inception of ICAD, and psychological human-subjects research has become as a legitimate approach to auditory display design and evaluation. However, little research has been conducted on modelling approaches to formalize human behavior in response to auditory displays. To bridge this gap, the present paper introduces computational modelling in auditory displays using the Queuing Network-Model Human Processor (QN-MHP) framework. After delineating the advantages of computational modelling and the QN-MHP framework, the paper introduces four case studies, which modelled drivers' behavior in response to in-vehicle auditory warnings, followed by the implications and future work. We hope that this paper can spark lively discussions on computational modelling in the ICAD community and thus, more researchers can benefit from using this method for future research.

1. INTRODUCTION

For over 25 years, the International Community for Auditory Display (ICAD) has developed and discussed a variety of methods and methodologies. Inherently, auditory display research is multidisciplinary and it involves different approaches, including art and design, as well as science and engineering. Given that the main goal of auditory displays is to make it easy for users to learn and use an interface, a human factors approach becomes a necessary method.

In human factors, different levels of approaches are required to validate a construct (e.g., [1]) just as in cognitive science [2]. On a psychological level, we can conduct experiments to show behavioral patterns. On a neurological level, we can use neuroimaging techniques (e.g., fMRI, fNIRS, EEG) to observe neural activities underlying behaviors. On a computational level, we can build mathematical models to simulate and predict behaviors. In the auditory display research community, psychological level approaches have seemed to settle down. Many papers have included a type of human subject tests, such as experiments, usability tests, focus groups or at least, interviews or questionnaires. When it comes to the neurological level, this dates back to Alvin Lucier's "Music for Solo Performer" using EEG data. Neuroimaging techniques have been used not just for neurological data-based sonification, but also for research on human behavior with music using fNIRS [3] or auditory cues using EEG [4]. In contrast, little research has been conducted on the computational level approaches in auditory displays.

The present study introduces the ICAD community to the use of computational modelling of behavioral responses to auditory displays. More specifically, we introduce four computational models of driver behaviors in response to in-vehicle auditory warnings in manual and automated vehicles with a focus on the Queueing Network-Model Human Processor (QN-MHP) framework. We hope that this paper will spark the lively discussion on the computational modelling approaches in the ICAD community, which will ultimately lead to a more balanced methodological practice for the development of auditory displays.

1.1 Why Computational Modelling?

Modelling is used in many domains in science and engineering. Models allow researchers to represent complicated phenomena as an abstract form and to simulate the dynamic state changes [5]. In sonification, design research is a necessary approach [6, 7], whereas modelling can also be a powerful research tool. Models are economical. Researchers do not always need to recruit participants to test different parameters on the outcomes. Models are accessible. Students and novices can easily modify the parameters to determine the effects of the changes without any design skills. *Computational* models have even more advantages. They quantify how different design parameters can influence human behaviors. Computational models enable simulations with diverse factors and predictions about future outcomes. They are often faster than real-time testing of large numbers of potential real-world scenarios, which may be difficult or even impossible to assess due to high risk or cost [8].

1.2 Why the QN-MHP Framework?

Different types of modelling approaches can be employed. Qualitative models explain human behaviors theoretically, whereas computational models provide quantitative relationships between the factors and the phenomena. Some models are built in a bottom-up fashion (e.g., machine learning), while others are built in a top-down fashion (e.g., cognitive architecture). Nowadays, much research has been conducted on modelling using the bottom-up approach, but little research has focused on human behavior modelling with respect to auditory displays. Likewise, in the traditional cognitive architecture approach, little research has been conducted with a focus on auditory displays. With the machine learning method, we can explore diverse unstructured data. However, it does not necessarily provide an explanation about the neurological or psychological mechanisms. In contrast, cognitive architecture can describe such mechanisms better. Among various cognitive architecture frameworks–e.g., ACT-R [9], EPIC [10],

SOAR [11], MHP, we focus on Queueing Network Model Human Processor (QN-MHP) [12] in the present paper.

The QN-MHP has many advantages to model human behaviors in response to auditory displays over other frameworks. QN-MHP has been used for modelling of response time based on neurological and psychological data. Its sub-network is built to represent a refined auditory system structure. It is one of the few frameworks that has been used for auditory display research. After briefly introducing the QN-MHP framework, we will delineate four case studies, all of which are in the automotive research area. QN-MHP is also good at simulating multitasking, which is often necessary for driving situations. Therefore, QN-MHP is appropriate for modelling drivers' behaviors in response to in-vehicle auditory displays.

1.3 What is the QN-MHP Framework?

The QN-MHP framework [12] represents the human cognition system as a queuing network based on several similarities to brain activities. QN-MHP consists of three subnetworks: perceptual, cognitive, and motor subnetworks, as described in Figure 2 (Appendix). Brain areas with similar functions are represented as servers and neural pathways connecting them as routes are represented as links in the queuing network, as seen in Figure 1.

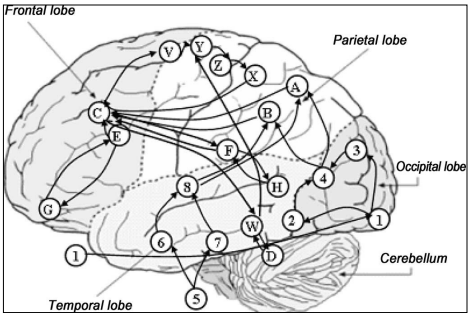


Figure 1: Approximate mapping of servers in the queuing Network Model onto the Human Brain [14].

A computational cognitive architecture called the queuing network-MHP (QN-MHP) has been developed as an integration of queuing networks and MHP for both mathematical modelling and real-time generation of psychological behavior. The queuing network model divides each process stage into a subnetwork of a small number of servers and thus, has a level of granularity that falls between the neural network and MHP models.

Each subnetwork is composed of multiple servers and links among these servers. Each server is an abstraction of a brain area with specific functions, and links among servers represent neural pathways among functional brain regions (Figure 2). The neurological processing of stimuli is illustrated in the transformation of entities passing through routes in QN-MHP.

QN-MHP has been successfully used to generate human performance and mental workload in real time, including driver performance and driver workload [14], transcription typing [15], and visual–manual tracking performance and mental workload measured by event-related potential (ERP) techniques. More basic information about QN-MHP is presented in [12] with specific explanations of its structure, assumptions, and implementation.

2. CASE STUDIES USING QN-MHP FRAMEWORK

In the present paper, we introduce four case studies to model drivers' behaviors in response to auditory displays in driving situations (see Table 1). The first paper modelled drivers' behaviors in response to speech warnings, the second paper to speech and non-speech warnings, and the last two papers improved modelling about non-speech warnings. We intentionally selected a line of research in the same domain to demonstrate how the models can be applied and enhanced to different situations (manual vehicles vs. automated vehicles), different auditory display types (speech vs. non-speech), and different modelling parameters (objective data vs. subjective data). Note that we did not explain the entire model of those studies, but we focused on auditory perception and processes.

2.1 Speech warnings in manual (connected) vehicles

The first study modelled driver behaviors in response to speech warnings in manual driving situations [16]. They modelled crash rate and response time.

2.1.1 Modelling

Modelling the effect of speech warning parameters on the probability of route choice

At server B (Phonological loop), there are two routes to move to either Server C (long route) or Server W (short route). The long route involves the activation of *accurate* hazard evaluation, whereas the short route involves a *rapid* automatic activation. The deciding probability of the two routes includes learning about the relationships between warning parameters and the hazard situations. Auditory stimuli are coactivated with the motor and premotor cortex (Server W) and the primary auditory cortex [17]. The probability of route choice is updated as participants learn from the association between loudness levels (or signal words) and hazard urgency.

The relationship between changes in loudness and changes in perceived urgency can be quantified by Stevens' Power Law [18]. Also, the loudness is positively correlated with urgency [18]. The perceived urgency (U_L) and annoyance (A_L) as a function of warning loudness were modelled as follows:

$$\log(U_L) = m_U \log(L) + k_U + \varepsilon 1 \quad (1)$$

$$\log(A_L) = m_A \log(L) + k_A + \varepsilon 2 \quad (2)$$

where L represents the loudness level and m and k quantify the relationship between perceived value and objective loudness change. The relationship between intensity and perceived urgency/annoyance was previously quantified [19]. Stevens' Power Law states that the loudness (L) is proportional to $I^{0.3}$, where I is the sound intensity [20]. Therefore, the parameters are quantified as: $m_U = 1.33$, $m_A = 1.45$, $k_U = -0.64$, $k_A = -0.91$. ε1 and ε2 are normally distributed random factors following distribution [0, 0.7] and [0, 0.86], respectively [21].

Table 1: Summary of each modelling work

	Zhang, 2016	Ko et al., 2019	Ko et al., 2019	Sanghavi, 2020
Cognitive Framework	QN-MHP	QN-MHP	QN-MHP	QN-MHP
Servers enhanced	6, 8, B, C, & F	8, B, C, & F	B, C, & F	8, B, C, & F
Independent variables of the experiment	Loudness and type of speech cues (warning, danger, caution), lead time	Type of auditory cues (speech, spearcons, earcons)	Type of non-speech cues (existing takeover sounds, indication sounds, and continuous sounds)	Driver emotions (anger and neutral), lead time, acoustic characteristics (fundamental frequency, number of repetitions of the sound)
Modelling parameters	Loudness, type of speech, lead time	Subjective rating results (perceived intuitiveness, perceived urgency)	Fundamental Frequency and number of repetitions of the sound, and the range of dominant frequencies	Fundamental Frequency and number of repetitions of the sound, and the range of dominant frequencies
Modelled variables	Crash rates, response time, and perceived urgency	Response time	Response time	Response time
Validation Results	Exp 1 Crash rate - RMSE: 0.13, R^2: 0.94 Response time - RMSE: 3.17, R^2: 0.97 Exp 2 Crash rate - RMSE: 0.06, R^2: 0.90 Perceived urgency - RMSE: 1.49, R^2: 1.00	RMSE: 0.073, R^2: 0.925 RMSE: 0.014, R^2: 0.999	RMSE: 1.48, R^2: 0.997	RMSE: 0.125, R^2: 0.505 RMSE: 0.119, R^2: 0.711

Modelling the effect of speech warning characteristics on the warning perception, memory decay and hazard evaluation

The effect of loudness on speech warning perception was modelled at Server 6 (Auditory Recognition) because the activation of the lower auditory processing level increased as the loudness increased [22]. The semantic features of signal words are recognized at the superior temporal sulcus, which was modelled at Server 8 (Auditory Recognition and location) [23].

The interference caused by the speech warnings on the ongoing tasks can cause memory decay [24]. Therefore, the effect of warning *lead time* on memory decay was modelled in the working memory system about auditory processing represented by Servers B (Phonological loop) and C (Central executive). fMRI studies showed that hazard evaluation activated the medial prefrontal cortex, the inferior frontal gyrus, the cerebellum, and the amygdala [25], which were presented by Server F (complex cognitive function).

The relationship between speech word choice and perceived urgency

Much research has shown a stable relationship between speech warning word choice and perceived urgency. Hollander and Wogalter [26] showed urgency ratings in a descending order: deadly, danger, warning, caution, and notice. The perceived urgency of "danger," "caution" and "notice" spoken by a female voice are quantified as 90.53,72.40 and 46.81 on a 100 points scale [18].

Speech warning parameters have different effects on response error rates in distinct stages of speech warning responses. When humans process speech warnings through route I, the error rate was mostly influenced by the effects of loudness and signal words on speech warning perception. However, when speech warnings are processed through route II, the error rate in the speech warning responses was also influenced by the effects of lead time on potential memory decay of the speech warnings and hazard evaluation. Thus, the two cases are modelled in different ways, with route I through server W directly (Appendix) described as the short route and route II through server C (Appendix) as the long route.

Modelling the response time to speech warnings

Given that response time is one of the most widely used dependent measures in cognitive psychology and human factors, it has been widely used in many modelling works. In QN-MHP, entity processing time at an individual server is independent of arrivals of entities, and routing is independent of the state of the system. In this case, the response time of a speech warning can be modelled by summing the processing time of all the servers on the route. Therefore, the response time (RTi) to speech warnings through route i is modelled as:

$$RT_i = T_5 + T_6 + T_8 + T_B + T_W + T_Y + T_Z,\ i = \text{I} \quad (3)$$

$$RT_i = T_5 + T_6 + T_8 + T_B + T_C + T_F + T_C + T_W + T_Y + T_Z,\ i = \text{II} \quad (4)$$

where T_k is the processing time of speech warning at Server k. The processing time of servers in perceptual,

cognitive, and motor subnetwork is 42ms, 24ms, and 18ms, respectively [13].

Equation (4) is updated by the effect of loudness on response time and the effect of word choice. The effect of loudness on response time is modelled in the initial processing of a speech warning in Server 6:

$$T_6 = T_{6(0)} / U_L \quad (5)$$

where $T_{6(0)}$ is the initial entity processing time in Server 6 and U_L denotes the effect of loudness on perceived urgency.

The effect of signal word choice on response time is modelled by the following equation:

$$T_8 = (T_{8(0)} \text{ X } n_i) / U_s \quad (6)$$

where $T_{8(0)}$ is the entity processing time in Server 8 and n_i is the number of words in the *i*th speech warning. U_s represents the urgency level expressed by the chosen words in the speech warnings.

2.1.2 Experiment and Validation

To validate this model, thirty-two participants (18 - 26 years old) participated in a driving simulation experiment. The speech warnings were generated before the appearance of the hazard. Each speech warning started with the signal word "Caution" and was followed by a description of the collision scenario presented (e.g., a vehicle at your front-left is running a red light) at 70 dBA with driving noise of 55 dBA. To investigate drivers' responses to speech warnings, their sights of the collision scenario were blocked by other vehicles (e.g., lead vehicles, parked vehicles or stacked vehicles at the intersection), and participants could only rely on the warnings to learn about the upcoming collision events. Each participant went through sixteen collision events with sixteen levels of lead time assigned to each event. The validation of the model was conducted via the Pearson correlation coefficient (R^2) and the root-mean-squared error (RMSE). For the crash rate with the speech warnings of different lead time levels, the model prediction comparing the experimental results had a RMSE of 0.13 with an R^2 of 0.94, which means the model was able to explain 94% of the data on average. For the brake-to-maximum response time to the speech warnings, the model prediction comparing the experimental results had a RMSE of 3.17 with an R^2 of 0.97.

The second experiment [21] investigated the effect of loudness and signal word choice of in-vehicle collision warnings. Thirty participants drove through five different scenarios containing five different hazard events. Speech warnings consisted of the signal word "Notice" or "Danger" presented at either 70 or 85 dBA. The model prediction for crash rate with different speech warnings comparing the experimental results had a RMSE of 0.06 with an R^2 of 0.90. The model prediction of rating of urgency and annoyance for signal word comparing the experimental results had a RMSE of 1.49 with an R^2 of perceived urgency prediction of 1.00. The R^2 of annoyance was not calculated as there is no difference among annoyance ratings of signal words [21].

2.2 Speech and non-speech takeover warnings in automated vehicles

The second study modelled driver behaviors in response to speech and non-speech takeover warnings in automated driving situations [27]. They modelled drivers' response time.

2.2.1 Modelling

The QN-MHP framework was further enhanced by accounting for the effect of warning reliability and warning style on human response to auditory warning messages [28, 29]. This was done specifically to include modelling human performance in speech warning responses and warning response type selection and execution. The warning response time for collision avoidance T_r (i, j) for a driver i in an event j, is modelled by the stimulus processing time of a route and the probability of a stimulus traveling through it, represented by the equation below [28, 29].

$$T_r(i,j) = \sum_{u=1}^{2} PT_u(i,j) \times P_u(i,j,wl,wsm,wt,ws,wr) \quad (7)$$

where *wl* and *wsm* represent the warning loudness and semantics respectively, *wt* is the warning lead time, *ws* represents the warning style and *wr* is the warning reliability. $PT_u(i, j)$ is defined as the processing time for a stimulus in driver i for an event j through a route u. This processing time through a route u was modelled by the addition of the total time of the stimulus running through all servers in the route u [16]. As seen in the previous section, there are two possible routes a signal can traverse for hazard perception as indicated by $u = 1$ (short) and $u = 2$ (long).

The probability of a speech warning traveling through a short processing route was modelled as a function of the perceived urgency of warning P_{wu}, perceived urgency of hazard P_{hu} and perceived trust in a warning P_{tr} [28, 29]. As it is shown in the equation below, the probability of traveling via a short route influenced by the perceived urgency of hazard P_{hu} is modelled as an inverse function of warning lead time *wt* [28, 29].

$$P_{hu} = 1/wt \quad (8)$$

The probability of traveling via a short route influenced by the perceived warning urgency P_{wu} is modelled as a function of warning loudness and warning semantics, where $U(wl)$ [30] is the perceived urgency of a warning signal's loudness and $U(wsm)$ is the perceived urgency of a warning signal's semantics [28, 29]. Also, the probability influenced by perceived trust of the warning P_{tr} was modelled as a function of warning reliability *wr* and warning style *ws* [28, 29]. The equations of P_{tr} and P_{wu} were described in the next section where a new model is developed based on this model. Ko et al. [27] enhanced this model by adding perceived intuitiveness and perceived urgency using subjective rating scales on different types of auditory warnings.

Trust is directly related to intuitiveness because trust requires adequate and intuitive communication [31] and heuristic processing tends to be primarily determined by belief in intuition [32]. The Rational versus Experiential Inventory (RVEI) [33] was used to measure propensity for rational processing and belief in the intuition scale indicating the extent to which participants use and rely on their intuition. Also, intuitiveness is related to accuracy and intelligibility since intuitive interaction is defined as a non-

challenging cognitive process in information-based activities [34]. In the previous studies on trust in the system, the increase in accuracy and intelligibility led to the increase in trust [35, 36]. It was assumed from these results that the intuitiveness of auditory cues is positively correlated with the perceived trust level of the cues. In the previous models [28, 29], the probability of warning information traveling via a short route influenced by the perceived trust of the warning P_{tr} was modelled by warning reliability *wr* and warning style *ws* as follows.

$$P_{tr} = 1/2\ (\mathrm{wr} + \sigma(\mathrm{ws}))\quad (9)$$

where σ(*ws*) is either 0.0 or 1.0 depending on the warning styles that are notification and command. The first variable of warning reliability was constant in Ko et al's study. The second variable of σ(*ws*) was replaced by *I(ws)*, which denotes the intuitiveness of the auditory warnings depending on the warning style. The value of *I(ws)* for speech was directly replaced by the maximum value for speech warning in the previous study [28, 29], which is 1.0, and other values were calculated depending on the relative differences between measured intuitiveness levels of speech and others in Experiment 1. Specifically, the values of *I(ws)* for spearcon and earcon are calculated by dividing the intuitiveness rating score of the auditory warning by the reference score (speech warning). As a result, P_{tr} was modified as

$$P_{tr} = 1/2\ (\mathrm{wr} + I(\mathrm{ws}))\quad (10)$$

where the value of *I(ws)* was estimated as 1.000 for speech, 0.825 for spearcon, and 1.05 for earcon.

The relationship between warning style and perceived urgency

In the previous models [28, 29], the probability of a warning traveling via a short processing route is influenced by perceived urgency of warnings P_{wu} , which was modelled as a function of warning loudness *wl* and warning semantics *wsm.*

$$P_{tr} = 1/2\ (U(wl) + U(wsm))\quad (11)$$

In Ko et al.'s study [27], the first variable of *U(wl)* was maintained since the loudness of sounds was controlled. *U(wsm)* was originally provided by the previous study on word choice [18], but was updated. To consider non-speech warnings such as earcon and spearcon, the variable of *U(wsm)* was replaced by *U(ws)*, which indicates the urgency level of auditory warnings. The value of *U(ws)* for speech was directly replaced by the value for speech warning in the previous model [28, 29], which is 90.53, and other values were calculated depending on the relative differences between measured urgency levels of speech and other sounds just as in the intuitiveness calculation. As a result, P_{wu} was modified as

$$P_{tr} = 1/2\ (U(wl) + U(ws))\quad (12)$$

where *U(ws)* is 90.530 for speech, 90.530 for spearcon, and 114.852 for earcon.

The response time in auditory warnings

The response time was defined as the time duration from the moment the auditory cue for takeover requests occurs to the moment the participant grabs the steering wheel. According to [16], the response time (T_r) of an auditory stimulus can be modelled by totalling the processing time of all the servers on the route as below.

$$T_r = (PT_5 + PT_6 + PT_8 + PT_B + PT_W + PT_Y + PT_Z)\ \mathrm{X}\ P_{\mathrm{I}} + (PT_5 + PT_6 + PT_8 + PT_B + PT_C + PT_F + PT_C + PT_W + PT_Y + \mathrm{P}T_Z)\ \mathrm{X}\ P_{\mathrm{II}}\quad (13)$$

where PT_{k} represents processing time at server *k* and P_k is the probability of choosing route *k* (*k* = I indicating the short route, and *k* = II indicating the long route). The probabilities of a stimulus traveling via the short route (*Route I*) can be calculated following the previous model [28, 29].

$$P_{\mathrm{I}} = 1/3\ (P_{wu} + P_{hu} + P_{tr})\quad (14)$$

$$P_{\mathrm{II}} = 1 - P_{\mathrm{I}}\quad (15)$$

where P_{wu}, P_{hu}, and P_{tr} are the probabilities influenced by perceived warning urgency, perceived hazard urgency, and perceived trust of the warning respectively.

Increased perceived urgency of an auditory warning leads to a decreased response time for a task performed [37, 38]. According to [16], the perceptual processing time is inversely affected by perceived urgency. The perceived urgency was modelled from the characteristics of warning sounds such as loudness and semantics. As mentioned, Hellier et al. [18] found that loudness had a positive relationship with perceived urgency.

In Ko et al.'s study [27], the processing time in the perceptual sub-network was calculated by using the directly measured perceived urgency level of each warning, *u*(*ws*). The value of *u*(*ws*) for speech was directly substituted by the value for speech warning in the previous study [28, 29], which is 1.0, and other values were calculated depending on the relative differences between measured urgency levels of speech and others. Specifically, the values of *u*(*ws*) for spearcon and earcon were calculated by dividing urgency rating score of the auditory warning by the reference score (speech). As a result, the equation of the response time of an auditory stimulus, Tr, was modified as

$$T_r = ((PT_5 + PT_6 + PT_8)/u(ws) + PT_B + PT_W + PT_Y + PT_Z)\ \mathrm{X}\ P_{\mathrm{I}} + (PT_5 + PT_6 + PT_8 + PT_B + PT_C + PT_F + PT_C + PT_W + PT_Y + \mathrm{P}T_Z)\ \mathrm{X}\ P_{\mathrm{II}} + \varepsilon\quad (16)$$

where ε is an added time as a free parameter which was estimated to the predicted processing time in the QN-MHP to fit their experimental data from Experiment 1 (ε = 1.060). The same value of this free parameter was used for validation with Experiment 2 data.

2.2.2 Experiment and Validation

To validate this model, twenty-two participants (*M* = 20 years old, *SD* = 1.7) participated in a driving simulation experiment. They chose three representative types of auditory warnings having unique characteristics and differences from each other. Specifically, speech was shown to be extremely learnable, whereas earcons were difficult to learn [39]. Male voice, "take over" was recorded for speech. For spearcons, the wave file of the speech clip was compressed by using the SOLA algorithm [40]. Sine wave

with two dominant frequencies (880, 1760 Hz) repeated four times was used for earcon following NHTSA [41] and ISO [42] guidelines. The amplitude and duration of all the sounds were controlled to be equivalent (Mean = 70dB, around 300ms). As a basic visual warning, the text, "Please take over" was displayed on the center monitor while auditory warnings were being generated. For the driving scenarios, participants were instructed to drive a car on a two-lane road in a rural area. Participants drove three laps in the experiment. A within-subjects experimental design with three auditory warning styles–speech, spearcon and earcon was applied in their experiment. The order of the conditions for each participant was counterbalanced. After a short driving, participants were instructed to take their hands off the steering wheel and their foot off the gas pedal. During the autonomous driving mode, the participants played the game, 2048 on a laptop placed next to them on the center console. After a while, the participants were triggered to take over from the autonomous driving mode by critical situations on the road such as deer, a parked car, and a service vehicle, that blocked most or all of the driving lanes on the road. A few seconds after avoiding the hazardous situations, the participants were instructed to take off the control and return to play the game again. Each participant continued driving for all three obstacles in the scenario and repeated the process in three conditions that generate different auditory warning styles.

First, the laps 2 and 3 data were modelled because the first lap data might reflect the significant learning effects. The prediction data showed the same pattern to the experimental data, which indicates the longest reaction time for spearcon and the shortest reaction time for earcon. The RMSE was 0.073 (73ms) with an R^2 of 0.925, which means the model was able to explain 92.5% of the experimental data on average.

When only using the final lap data, the prediction data also showed the same pattern in the experimental data, which indicates the longest reaction time for spearcon and the shortest reaction time for earcon. The RMSE was 0.014 (14ms) with an R^2 of 0.999. Both RMSE and R^2 values in validation 2 were higher than those values in validation 1.

2.3 Non-speech takeover warnings in automated vehicles

The next study modelled driver behaviors in response to non-speech takeover warnings in automated driving situations [27]. They modelled drivers' response time by varying types of non-speech auditory cues.

2.3.1 Modelling

The relationship between warning style and perceived trust

In this modelling work, three different types of non-speech sounds (two already implemented sounds and one newly designed sound) were employed. Perceived urgency that influences the probability of choosing a short route, P_{wu}, was estimated based on acoustic characteristics–fundamental frequency (Hz), number of repetitions/second, and number of dominant frequencies. A new parameter, $U(wac)$, was added to represent the perceived urgency from acoustic characteristics of auditory cues. P_{wu} was modelled as below.

$$P_{wu} = 1/3\ (U(wl) + U(wsm) + U(wac)) \quad (17)$$

The values of $U(wac)$ was estimated as below.

$$U(wac) = 1/3\ (U(wfreq) + U(wrep) + U(wpit)) \quad (18)$$

U(wfreq) represents the fundamental frequency, *U(wrep)* represents the number of repetitions per second, and *U(wpit)* represents the pitch range of dominant frequencies. Coefficients in each term were derived from the previous studies (fundamental frequency [43], number of repetitions, and pitch range: [44]) showing linear relationships between each aspect and perceived urgency.

2.3.2 Experiment and Validation

Twenty-four participants (M = 20 years old, SD = 1.1) participated in the experiment. Twelve participants' data were used for modelling. The procedure of the experiment was exactly same as the experiment reported in the previous section except for the sounds used. The experiment included existing takeover warnings (Tesla and Hyundai Motors), a pair of indication sounds that can be easily heard from electronic devices (e.g., function on and function off sounds that have increasing and decreasing polarity respectively). The last pair of sounds included rather longer musical sounds that had one or two fundamental frequencies. But the average time was similar to that of the short sounds (by repeating them).

In modelling, Tesla, Hyundai Motors, and indication sounds with decreasing polarity were used. The average response times of each auditory display type for takeover requests were predicted and validated with the validation data sets from the remaining twelve participants. The RMSE was 0.148 (148 ms) with an R^2 of 0.997, which means the model was able to explain 99.7% of the experimental data on average. In both prediction and validation data sets, the longest reaction time was observed for Tesla sound and the shortest reaction time for Hyundai Motors sound.

2.4 Non-speech takeover warnings for drivers in automated vehicles

The final study modelled driver behaviors in response to non-speech takeover warnings in automated driving situations [45]. They modelled drivers' response time by varying the acoustic variables of earcons.

2.4.1 Modelling

This study used the same modelling scheme with acoustic parameters as the third study.

$$U(wfreq) = 12 \times \left(F_{fundamental} - 210\right)/470 + 67.2 \quad (19)$$

$$U(wrep) = \left(30.2295 \times log_{10}[N_{repetitions\ per\ second}]\right) + 45 \quad (20)$$

$$U(wpit) = \left(1.8553 \times D_{pitch}\right) + 42.6053 \quad (21)$$

2.3.2 Experiment and Validation

Thirty-six participants took part in the study (M = 24.5 years old, SD = 3.3). Each participant had a minimum of two years of driving experience and was at least 18 years old. Participants were randomly assigned to two groups: an anger induced group and a neutral induced group. However, there was no difference in their response time to takeover

displays. Therefore, the validation was conducted with the aggregated data of both emotional groups.

The participants were asked to drive a semi-automated vehicle on the center lane of a three-lane highway. The vehicle was moving at a speed of 110 km/h. The visibility around the vehicle was reduced to 100 meters using fog. The scenario was started with the vehicle having its automated driving system (ADS) turned on. At randomly assigned times, the participants were required to respond to different multi-modal displays (an auditory plus visual display on the screen showing the words, "Takeover Control of the Vehicle" during takeover requests) in specific driving scenarios where drivers had to evade obstacles, with the auditory warnings differing for each obstacle. After avoiding the obstacles, the participants handed over control back to the vehicle. Each participant drove a total of three laps which lasted around nine minutes. Every lap had three obstacles each and so, each participant faced nine different obstacles (3 x 3). The order of alert presentation was counterbalanced with a Latin square design.

The RMSE was 0.125. The R^2 for the model was 0.505 for all drivers combined, which means the model was able to explain 50.5% of the data on average. A higher number of repetitions in auditory warnings would result in higher perceived urgency and thus, faster reaction times. This increase in perceived urgency due to the increase in the number of repetitions per second was incorporated into the model as well.

However, analysis from the empirical data showed that although an increase of one repetition per second to four repetitions per second resulted in faster reaction times, the increase to eight repetitions per second did not show a decrease in reaction time. Thus, the low correlation of the data with the modelled predictions was because the reaction time for the eight repetitions auditory warnings did not show a decrease as the model predicted it should be. After removing reaction times for eight repetitions per second for the correlation analysis, the RMSE of 0.119 with the R^2 being significantly improved to 0.711, representing a 20.6% increase in the data explained by the model on average.

3. DISCUSSION AND FUTURE WORK

We showed four different modelling and validation case studies using QN-MHP to describe the applications and enhancements of computational modelling with respect to auditory displays. The studies included speech and non-speech cues. Except for the first study, all the studies modelled drivers' takeover response time to auditory warnings in semi-automated vehicles. With the same architecture, they advanced the model using a variety of parameters, including objective and subjective data– loudness, word of speech, perceived intuitiveness and urgency, fundamental frequency, number of repetitions, and range of dominant frequencies. The first and last studies also included lead time, which was not detailed in the present paper. These models discussed the effects of warning characteristics on auditory perception, auditory recognition, route choice and memory decay in the working memory (Phonological loop and Central executive), and hazard evaluation, following the neurological mechanisms across the perceptual, cognitive, and motor subnetworks. The models showed robust and reliable prediction values (mostly, over 90%). The last study used the same modelling scheme as the third one even though the experimental procedure was not exactly same. Thus, it showed the applicability of the modelling to a different situation. The model could still account for 71% of the experimental data, which is promising.

We demonstrated that the QN-MHP framework is one tool we can use, but different modelling techniques are available and how these techniques can be made more relevant for the ICAD community should be tested. This task will include comparisons of different approaches–e.g., machine learning (bottom-up) vs. top-down (cognitive architecture informed), vs. hybrid models. Modelling multimodal interactions, beyond audio-only stimuli would be of interest. We can further discuss the way modelling approaches can be aligned with other approaches (e.g., user studies, quantitative experiments, designs, prototyping) so that they can complement and strengthen each other.

As with other methods, computational modelling has limitations. In some cases, free parameters are inserted to modify the equation to fit the empirical data. It is hard to pinpoint where the gaps occur. Even though the approach described here is based on known neurological and psychological knowledge and mechanisms, there might be incorrect assumptions. For instance, it assumes all the processes in each subnetwork and server happen independently and so, the response time is obtained by summing sequential temporal periods. This is a fundamental assumption in cognitive psychology, but there is still argument about it. It is also hard to consider the integration procedure of various information pieces. This type of modelling work is based on ideal responses to each stimulus of the sampled group (e.g., young adults). Therefore, the prediction may not work for other groups of people (e.g., older adults). Finally, compared to psychological experiments, few researchers have used the computational modelling approach. Consequently, more cumulative research is required.

In summary, with computational modelling, researchers can more systematically investigate the quantitative relationship between sound parameters and human perceptions and behaviors by manipulating speech and acoustic parameters as can be seen in Table 1. Computational modelling can save time and money because it can provide estimated behavioral outcomes without conducting empirical experiments with human participants. The systematic results can guide designers to design appropriate auditory displays in the given contexts and constraints.

REFERENCES

[1] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs," Journal of cognitive engineering and decision making, vol. 2, no. 2, pp. 140-160, 2008.

[2] D. Marr, "Vision: A computational investigation into the human representation and processing of visual information," 1982.

[3] V. Straebel, and W. Thoben, "Alvin Lucier's music for solo performer: experimental music beyond sonification," 2014.

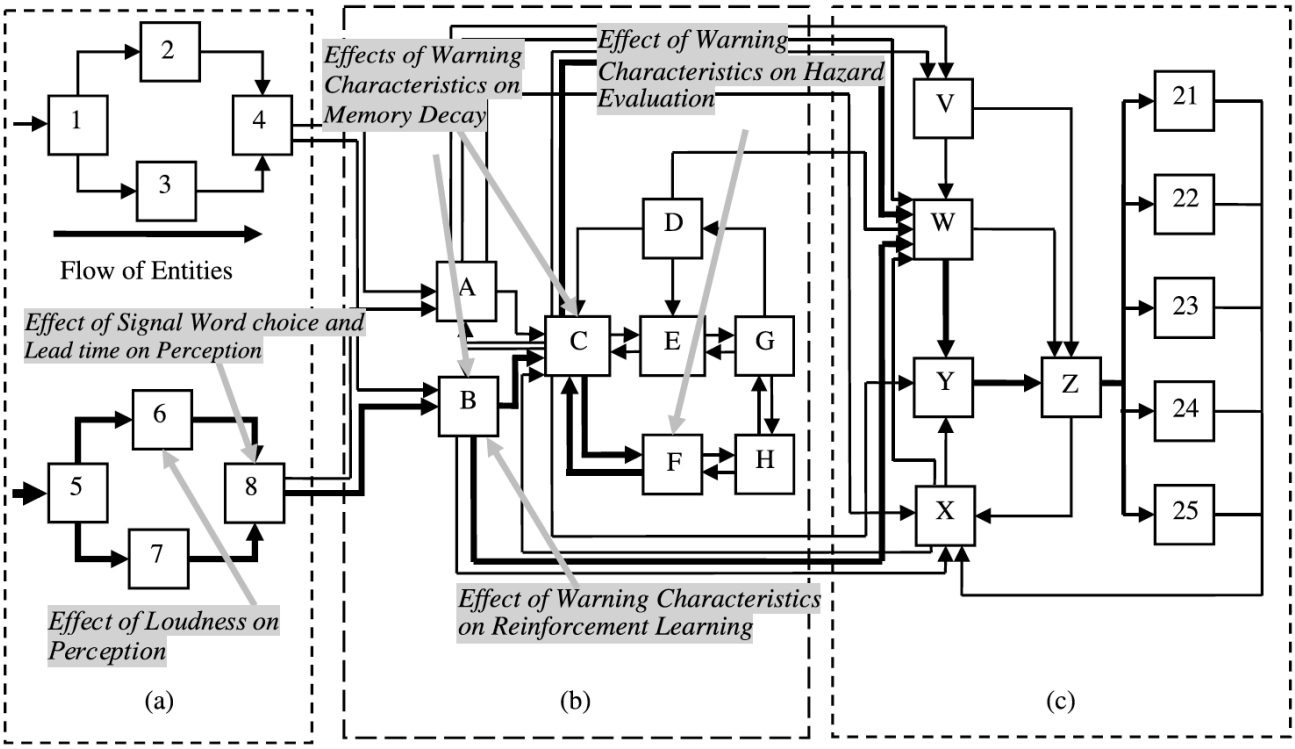
[4] L. L. Chuang, C. Glatz, and S. Krupenia, "Using EEG to understand why behavior to auditory in-vehicle notifications differs across test environments." AutoUI 2017. pp. 123-133. 2017

[5] J. L. McClelland, "The place of modeling in cognitive science," Topics in Cognitive Science, vol. 1, no. 1, pp. 11-38, 2009.

[6] M. Jeon, B. N. Walker, and S. Barrass, "Introduction to the special issue on sonic information design: Theory, methods, and practice, part 2," Ergonomics in Design, vol. 27, no. 1, p.4, 2019.

[7] M. Jeon, B. N. Walker, and S. Barrass, "Introduction to the special issue on sonic information design: Theory, methods, and practice, part 1," Ergonomics in Design, vol. 26, no. 4, p.3, 2019.
[8] A. D. McDonald, H. Alambeigi, J. Engström, G. Markkula, T. Vogelpohl, J. Dunne, and N. Yuma, "Toward computational simulations of behavior during automated driving takeovers: a review of the empirical and modeling literatures," Human Factors, vol. 61, no. 4, pp. 642-688, 2019.
[9] J. R. Anderson, "ACT: A simple theory of complex cognition," American psychologist, vol. 51, no. 4, pp. 355, 1996.
[10] D. E. Kieras, and D. E. Meyer, The EPIC architecture for modeling human information-processing and performance: A brief introduction, Michigan Univ Ann Arbor Div Of Research Development and Administration, 1994.
[11] J. F. Lehman, J. E. Laird, and P. Rosenbloom, "A gentle introduction to Soar, an architecture for human cognition," Invitation to cognitive science, vol. 4, pp. 212-249, 1996.
[12] Y. Liu, R. Feyen, and O. Tsimhoni, "Queueing Network-Model Human Processor (QN-MHP) A computational architecture for multitask performance in human-machine systems," ACM Transactions on Computer-Human Interaction (TOCHI), vol. 13, no. 1, pp. 37-70, 2006.
[13] C. Wu, and Y. Liu, "Queuing network modeling of the psychological refractory period (PRP)," Psychological Review, vol. 115, no. 4, pp. 913, 2008.
[14] C. Wu, and Y. Liu, "Queuing network modeling of driver workload and performance," IEEE Transactions on Intelligent Transportation Systems, vol. 8, no. 3, pp. 528-537, 2007.
[15] C. Wu, and Y. Liu, "Queuing network modeling of transcription typing," ACM Transactions on Computer-Human Interaction (TOCHI), vol. 15, no. 1, pp. 1-45, 2008.
[16] Y. Zhang, C. Wu, and J. Wan, "Mathematical modeling of the effects of speech warning characteristics on human performance and its application in transportation cyberphysical systems," IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 11, pp. 3062-3074, 2016.
[17] F. Pulvermüller, M. Härle, and F. Hummel, "Walking or talking?: Behavioral and neurophysiological correlates of action verb processing," Brain and language, vol. 78, no. 2, pp. 143-168, 2001.
[18] E. Hellier, J. Edworthy, B. Weedon, K. Walters, and A. Adams, "The perceived urgency of speech warnings: Semantics versus acoustics," Human Factors, vol. 44, no. 1, pp. 1-17, 2002.
[19] C. Gonzalez, B. A. Lewis, D. M. Roberts, S. M. Pratt, and C. L. Baldwin, "Perceived urgency and annoyance of auditory alerts in a driving context." Proceedings of the Human Factors and Ergonomics Society Annual Meeting. pp. 1684-1687.
[20] J. Parmanen, "A-weighted sound pressure level as a loudness/annoyance indicator for environmental sounds–Could it be improved?," Applied Acoustics, vol. 68, no. 1, pp. 58-70, 2007.
[21] C. L. Baldwin, "Verbal collision avoidance messages during simulated driving: perceived urgency, alerting effectiveness and annoyance," Ergonomics, vol. 54, no. 4, pp. 328-337, 2011.
[22] I. S. Sigalovsky, and J. R. Melcher, "Effects of sound level on fMRI activation in human brainstem, thalamic and cortical centers," Hearing Research, vol. 215, no. 1-2, pp. 67-76, 2006.
[23] S. Uppenkamp, I. S. Johnsrude, D. Norris, W. Marslen-Wilson, and R. D. Patterson, "Locating the initial stages of speech–sound processing in human temporal cortex," Neuroimage, vol. 31, no. 3, pp. 1284-1296, 2006.
[24] K. R. Laughery, "Computer simulation of short-term memory: A component-decay model," Psychology of Learning and Motivation, pp. 135-200: Elsevier, 1970.
[25] V. Vorhold, C. Giessing, P. Wiedemann, H. Schütz, S. Gauggel, and G. Fink, "The neural basis of risk ratings: Evidence from a functional magnetic resonance imaging (fMRI) study," Neuropsychologia, vol. 45, no. 14, pp. 3242-3250, 2007.
[26] T. D. Hollander, and M. S. Wogalter, "Connoted hazard of voiced warning signal words: an examination of auditory components." Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 44, No. 22, pp. 702-705), 2000.
[27] S. Ko, Y. Zhang, and M. Jeon, "Modeling the effects of auditory display takeover requests on drivers' behavior in autonomous vehicles." AutoUI 2019, pp. 392-398, 2019.
[28] Y. Zhang, and C. Wu, "Modeling the Effects of Warning Lead Time, Warning Reliability and Warning Style on Human Performance Under Connected Vehicle Settings." Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 62, No. 1 pp. 701-701), 2018.
[29] Y. Zhang, "Mathematical Modeling of Driver Performance in Warning Responses under the Connected Vehicle Settings," PhD Dissertation. University at Buffalo, 2017.
[30] S. Nabe, S. J. Cowley, T. Kanda, K. Hiraki, H. Ishiguro, and N. Hagita, "Robots as social mediators: coding for engineers." The 15th IEEE International Symposium on Robot and Human Interactive Communication, pp. 384-390. 2006.
[31] C. D. Wickens, "Multiple resources and performance prediction," Theoretical issues in ergonomics science, vol. 3, no. 2, pp. 159-177, 2002.
[32] D. A. Krauss, J. G. McCabe, and J. D. Lieberman, "Dangerously misunderstood: Representative jurors' reactions to expert testimony on future dangerousness in a sexually violent predator trial," Psychology, Public Policy, and Law, vol. 18, no. 1, pp. 18, 2012.
[33] S. Epstein, R. Pacini, V. Denes-Raj, and H. Heier, "Individual differences in intuitive–experiential and analytical–rational thinking styles," Journal of Personality and Social Psychology, vol. 71, no. 2, pp. 390, 1996.
[34] J. Hurtienne, C. Mohs, H. A. Meyer, M. C. Kindsmüller, and J. Habakuk Israel, "Intuitive use of user interfaces-definition und herausforderungen," i-com vol. 5, no. 3 pp. 38-41. 2006.
[35] P. W. Bonsall, and M. Joint, "Driver compliance with route guidance advice: the evidence and its implications." In Vehicle Navigation and Information Systems Conference, vol. 2, pp. 47-59. 1991.
[36] J. E. Fox, and D. A. Boehm-Davis, "Effects of age and congestion information accuracy of advanced traveler information systems on user trust and compliance," Transportation Research Record, vol. 1621, no. 1, pp. 43-49, 1998.
[37] E. C. Haas, and J. G. Casali, "Perceived urgency of and response time to multi-tone and frequency-modulated warning signals in broadband noise," Ergonomics, vol. 38, no. 11, pp. 2313-2326, 1995.
[38] E. C. Haas, and J. Edworthy, "Designing urgency into auditory warnings using pitch, speed and loudness," Computing & Control Engineering Journal, vol. 7, no. 4, pp. 193-198, 1996.
[39] B. N. Walker, J. Lindsay, A. Nance, Y. Nakano, D. K. Palladino, T. Dingler, and M. Jeon, "Spearcons (speech-based earcons) improve navigation performance in advanced auditory menus," Human Factors, vol. 55, no. 1, pp. 157-182, 2013.
[40] S. Roucos, and A. Wilgus, "High quality time-scale modification for speech." In ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 10, pp. 493-496. 1985.
[41] J. Campbell, J. Brown, J. Graving, C. Richard, M. Lichty, T. Sanquist, and J. Morgan, "Human factors design guidance for driver-vehicle interfaces," National Highway Traffic Safety Administration, Washington, DC, DOT HS, vol. 812, pp. 360, 2016.
[42] ISO., "ISO Standard 15006:2011(E). In Road vehicles—Ergonomic aspects of transport information and control systems—Specifications for in-vehicle auditory presentation: international organisation for standardisation.," (ISO), 2011.
[43] C. L. Baldwin, and B. A. Lewis, "Perceived urgency mapping across modalities within a driving context," Applied ergonomics, vol. 45, no. 5, pp. 1270-1277, 2014.
[44] J. Edworthy, S. Loxley, and I. Dennis, "Improving auditory warning design: Relationship between warning sound parameters and perceived urgency," Human factors, vol. 33, no. 2, pp. 205-231, 1991.
[45] H. Sanghavi, "Measuring the influence of anger on takeover performance in semi-automated vehicles." Virginia Polytechnic Institute and State University, Blacksburg, VA., 2020.

Appendix.



(a) Perceptual Subnetwork (b) Cognitive Subnetwork (c) Motor Subnetwork

1. Common visual processing
2. Visual recognition
3. Visual location
4. Visual recognition and location integration
5. Common auditory processing
6. Auditory recognition
7. Auditory location
8. Auditory recognition and location

A. Visuospatial sketchpad
B. Phonological loop
C. Central executive
D. Long-term procedural memory
E. Performance monitor
F. Complex cognitive function
G. Goal initiation
H. Long-term declarative & spatial memory

V. Sensorimotor integration
W. Motor program retrieval
X. Feedback information collection
Y. Motor program assembling and error detecting
Z. Sending information to body parts
21-25: Body parts: eye, mouth, left hand, right hand, foot

Figure 2: The general structure of the queuing network-model human processor adapted from [16]. The short route goes from server B to server W, while the long route goes from server B to server C.