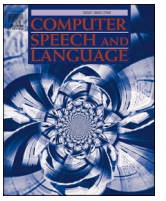
Computer Speech and Language 83 (2023) 101537

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl



# The use of Active Learning systems for stimulus selection and response modelling in perception experiments

Marieke Einfeldt [a,*], Rita Sevastjanova [b], Katharina Zahner-Ritter [c], Ekaterina Kazak [d], Bettina Braun [a]

[a] Department of Linguistics, University of Konstanz, Germany
[b] Department of Computer and Information Science, University of Konstanz, Germany
[c] Phonetics, University of Trier, Germany
[d] Economics/School of Social Sciences, University of Manchester, United Kingdom

ARTICLE INFO

*Keywords:*
Psycholinguistics
Cue weighting research
Prosody
Active learning
Stimulus selection
Modelling
Linguistic architecture

ABSTRACT

To study the role of perceptual cues on categorization and decision making, participants are typically tested in (perception) experiments with a fixed set of randomized or pseudo-randomized trials. In linguistics and psycholinguistics, for instance, studies often investigate the relative weighting of different cues for a linguistic contrast (e.g., intonation vs. word order). For categorization beyond the segmental level (e.g., /p/ vs. /b/), it is important to establish that results generalise to different words or sentences, which necessitates the use of a range of different items. This may limit the number of conditions (cues and cue combinations) that can be sensibly tested in the same experiment. We show that Active Learning (AL) systems provide a solution: Since stimulus selection is informed by the system's learning mechanism (presenting certain conditions less often than uncertain conditions), they allow for efficient testing of numerous conditions and different items in the same experiment. In this paper, we compared two weighting approaches (average probability-based vs. regression-based) to model the outcome of three simulated scenarios with three binary factors each. Results show that valid results (i.e., little error between predicted values and the actual responses at the end of the experiment) are obtained after about a third of the trials of an original psycholinguistic experiment we replicated. For simulations with interactions between factors, the regression-based approach performed better. Our findings bear implications for the application of AL in psycholinguistic research (extraction of cue weights, inferential statistics, and a stopping criterion during an on-going experiment), which we will discuss.

## 1. Introduction

Linguistic contrasts are often signalled by a number of different cues. For example, polar (yes/no) questions and declarative statements differ in terms of their syntactic structure (in English verb-first interrogative compared to verb-second declarative sentences) and intonational realization (high rising boundary tones in polar questions compared to falling ones in declaratives, Ladd, 2008). There are further, more subtle cues, such as segment durations and the nature of intonational realization at the start of the

* Corresponding author.
*E-mail address:* marieke.einfeldt@uni-konstanz.de (M. Einfeldt).

https://doi.org/10.1016/j.csl.2023.101537
Received 1 September 2022; Received in revised form 17 May 2023; Accepted 24 June 2023
Available online 26 June 2023

utterance that distinguish the meaning (e.g., Petrone and Niebuhr, 2014). Studying the contribution of all these individual cues and their relative importance is essential for our understanding of language processing (Schertz and Clare, 2020; Stevens and Klatt, 1974) as well as linguistic theory and architecture. Linguistic modelling needs to be informed on the role of the cues: whether they are redundant with equal weight, such that one cue is sufficient to signal the contrast, whether each has their fixed contribution, with similar or different cue weights, or whether one may be traded against another.

Cue weighting studies have been conducted in a variety of different linguistic areas, e.g., for lexical stress (e.g., Chrabaszcz et al., 2014; Frost, 2011; Fry, 1958; Gordon and Roettger, 2017; Kaland, 2020; Kohler, 2008, 2012; van Heuven and de Jonge, 2011; Zahner et al., 2019), phrasal prominence and information structure (e.g., Andreeva et al., 2007, 2012; Barry and Andreeva, 2011; Baumann, 2014; Baumann and Röhr, 2015; Baumann and Winter, 2018; Cole et al., 2010; Cole and Shattuck-Hufnagel, 2016; Wagner et al., 2016), prosodic boundaries (e.g., Beach, 1991; Lehiste et al., 1976; Mo and Cole, 2010; Petrone et al., 2017; Zhang, 2012), sentence modality (e.g., Cangemi and D'Imperio, 2013; Genzel and Kügler, 2020; Shiamizadeh et al., 2017), case ambiguity (e.g., Gollrad et al., 2010) and pragmatic functions, such as irony or rhetorical illocution (e.g., Braun and Schmiedel, 2018; Cheang and Pell, 2008; Kharaman et al., 2019; Nauke and Braun, 2011).

For studies at the prosody-semantics interface, a large number of cues may be potentially relevant and hence need to be tested in an orthogonal design to allow for valid conclusions. At the same time, different lexicalisations/sentences need to be employed in order to ensure generalisability (cf. Cutler, 1977, who already discussed the context-dependency of intonational meaning). For instance, depending on your taste or moral, the question "Who eats meat?" may be more readily interpreted as a rhetorical question (similar to a statement that nobody eats meat), a biased question (in which you imply that you expect others to not eat meat), or an information-seeking question. By exchanging objects and the related predicates, i.e., the lexicalisations/sentences, experimenters try to mitigate potential (individual) interpretation. Combining both design desiderata (large set of orthogonally manipulated cues and different lexicalisations) in a traditional experiment in which a participant rates each experimental stimulus would result in a very high number of trials.[1] Conducting an experiment with too many trials has at least three downsides: (a) the participants may get bored, exhausted or entrained and the results might therefore be hard to generalize, (b) the experiment is time-consuming for the participants, which reduces their willingness to participate in the first place, and (c) it is expensive for the experimenters. Theoretically, researchers can choose between one of three options to reduce the time invest of participants to complete an experiment and mitigate those downsides: (i) a larger number of conditions (or cues) is compensated by a smaller number of lexical items (down to $N = 1$), (ii) a larger number of lexical items is compensated by reducing the number of conditions or (iii) a high number of conditions and lexical items is distributed over several experiments with different participants. In practice, (i) is most frequently encountered (e.g., Frost, 2011; Genzel and Kügler, 2020; Kohler, 1991, 2012; Niebuhr and Winkler, 2017; van Heuven and de Jonge, 2011). For instance, van Heuven and de Jonge (2011) investigated the relative roles of spectral and durational cues in stress perception in Dutch, crossing duration and spectral properties in a 7×7 design (i.e., 49 conditions) on a single minimal stress pair (cf. Kohler, 2012, for a similar design on German). While such studies provide a very fine-grained picture on the weighting and interplay between individual cues, generalization to other items (e.g., with a different number of syllables, different syllable structures or different lexical frequencies of the members of a pair) is limited. Studies that use a high number of lexicalisations and fewer cues (Kharaman et al., 2019; Zahner et al., 2019), by contrast, ensure a higher generalizability, but are restricted as to the number of cues/conditions that can be tested. Kharaman et al. (2019), for instance, the study we aim to replicate in the present paper using Active Learning (AL), examined the relative contributions of pitch accent type (2 levels), duration (2 levels), and voice quality (2 levels) for the interpretation of an utterance as an information-seeking or rhetorical question. This 2×2×2 design resulted in 8 conditions, which were distributed over 32 lexical items (the cues were manipulated partly within-items, partly between-items).

It is clear that solutions are needed to satisfy the need of testing multiple conditions with multiple lexicalisations and at the same time keeping the experiment feasible. Active Learning (AL) presents one tool or solution. In the present study, we present three simulated scenarios with a 2×2×2 design with different outcomes and test the validity of prediction and the speed of a valid prediction with two different weighting algorithms.

### 1.1. General advantages of active learning systems

AL is a sub-field of machine learning in which a learning algorithm queries an oracle (e.g., a human annotator) to annotate unlabelled stimuli with pre-defined class labels (Settles, 2009) and derives rules for the labelling of similar stimuli. AL techniques have been employed in computer science research since the 1980s (Angluni, 1988); they have been used, amongst others, for named-entity recognition (Shen et al., 2004), semantic parsing (Thompson et al., 1999), and text classification (McCallum and Nigam, 1998). Using AL for cue weighting research necessitates a rethinking of the role of the annotator and the output: In our use case, the annotator is the experimental participant and the output is not the actual label, but the cue weights that are obtained from the AL model over the course of the experiment.

In short, AL systems apply algorithms to query label information from a user, with the selection of stimuli being based on an optimisation of the classifier's performance (Settles, 2009), i.e., an improvement of its prediction accuracy. AL optimizes the order in which the instances/stimuli are labelled by applying an appropriate sampling strategy. Different sampling strategies have been

[1] For instance, a simple setup with three binary cues that are tested on sixteen sentences (manipulated within-items, so that each sentence is presented with each combination of three cues) already results in 128 stimuli. If one of the cues is not binary but has multiple steps (e.g., 5), this number rises to 640.

presented in the literature, commonly being classified as *data-centred* or *model-based strategies*. Data-centred strategies rely on the characteristics of the corpus and query labels for instances according to their similarity or density (Wu et al., 2006). Model-based strategies, in contrast, rely on the suggestions of an underlying machine learning model trained on iteratively labelled data instances. Using different criteria, such as error reduction (Settles, 2012), classifier uncertainty (Smallest Margin, Wu et al., 2006), or entropy (Vendrig et al., 2002), the system asks the user to label instances that improve the model's performance most with respect to a specific quality metric (e.g., model's accuracy, Japkowicz and Shah, 2015). The underlying model can be of various complexity, ranging from simple Association Rule Mining models or Linear Regression models to more complex Support Vector Machines or even Deep Learning models. These sampling strategies make the labelling process more efficient and, hence, reduce the human effort needed to obtain a labelled corpus.

In this paper, we use model-driven uncertainty-based sampling to select stimuli for labelling, which is one of the most popular approaches in AL (Nguyen et al., 2022), see Fig. 1. Since the AL system iteratively learns a probabilistic classification model on the basis of the labelled stimuli, one can use this model to predict class labels for unlabelled stimuli, which is useful for very complex designs (cf. Braun et al., 2022). This generally allows researchers to restrain the number of trials needed for the generation of a labelled corpus or experimental data and, hence, reduces the human effort. As mentioned before, the AL system can utilize different classification models with varying degrees of complexity. Since the entry point of our experiments is a fully unlabelled corpus, we use two kinds of simple classification models (average probability-based approach and regression-based approach) that are both able to predict class labels by learning from only a small number of labelled stimuli. They differ in their properties, which are described in more detail in Sections 4 and 5. In short, the average probability-based approach predicts the class label by averaging probability values of its observed components (i.e., cue combinations), whereby each component is equally weighted.

In contrast, the regression-based approach replaces the equal weighting of the components by the weights, which are estimated from the data. Both average probability-based approach and regression-based approach update the probability that a given cue combination belongs to a certain category in a data-driven way. However, the contribution of different cues to that probability for the average probability-based approach is fixed to be the equal positive weight $1/n$, where $n$ is the total number of cue combinations included to the analysis. The regression-based approach allows for more flexible contributions of different cue combinations, including negative weights. Since one does not know the outcome of cue weights a-priori, we compared these two methods to investigate which is more valid in different scenarios.

Previous studies with AL have shown that rules can be derived from a small set of labelled instances (Sevastjanova et al., 2018) and classifications can be validated by presenting similar conditions with other items. We see three main benefits for using AL in cue weighting research:

(1) AL allows for the testing of many conditions with fewer trials, as stimulus selection is informed by the system's learning mechanism. In the classical experiments, which do not use AL, we implicitly assume that we need the same number of trials per cue combination to estimate the effect of this combination. Using AL allows us to select the cue combinations for labelling proportional to the uncertainty of the response. Therefore, the effective sample size for more uncertain cue combinations will be larger and thus allow for more precise estimates of cue effects.
(2) The cue weights, i.e., the importance of each cue, can be estimated from the participants' responses. The weights of cues and the combination of individual cues that the system calculates also help to model the relationships between cues.
(3) The data can still be analysed with classical methods of inferential statistics to corroborate the interpretation.

More detail on the AL approach is provided in Sections 4 and 5, with actual examples from the prediction task to make the modelling more accessible.

### 1.2. Aims and structure of this paper

The present paper has two aims: First, we compare the predictions of the AL systems for three scenarios. One with simple main effects for the three cues (Scenario I), one with a two-way interaction (Scenario II) and one with a three-way-interaction (Scenario III).

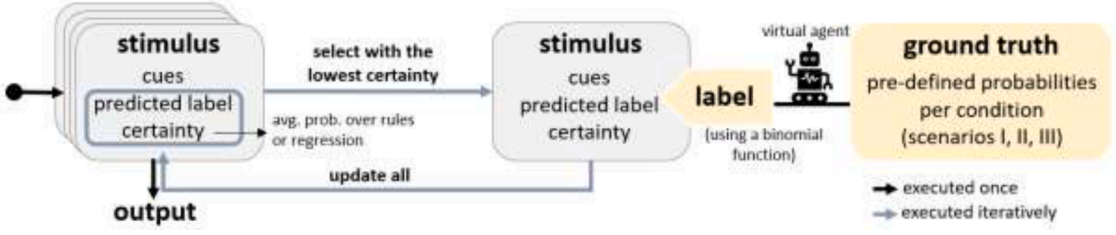



**Fig. 1.** In our AL approach, the data (stimuli and their cues) is first read into the system. The labelling approach begins by randomly selecting a stimulus from the dataset, which is then labelled by the virtual agent using a binomial function on the pre-defined probabilities per condition. Based on the provided label and the cues of the labelled stimulus, the predicted label, as well as the prediction certainty, get updated for all stimuli in the dataset. The next stimulus for labelling is the one with the lowest certainty. The process is repeated until satisfying a pre-defined stopping criterion.

In the two scenarios with interactions, the effect of one cue depends on the level of the other cue which may need more data to model the cue weights accurately. For evaluation, we compare the root-mean-square error (RMSE) between the responses at each trial and a gold standard. As gold standard we use the actual responses at the last trial of the experiment (64 trials). A low RMSE indicates a high validity. The earlier in the experiment this can be reached, the better. To evaluate the speed with which a stable prediction is achieved, we extracted the trial at which there are five consecutive trials with less than 5% reduction/improvement in RMSE, relative to the average RMSE of the preceding five trials. This measure is secondary as an early stabilisation at a high RMSE is not desirable. Further, we compare two different approaches for the learning algorithm, average probability-based approach and regression-based approach, which have different benefits. For reasons of efficiency, the actual responses are collected by virtual agents instead of actual participants (except for the experiment presented in Section 2). These virtual agents are programmed on pre-defined probabilities per condition using a binomial sampling function.

The paper is structured as follows: Section 2 presents the behavioural experiment that is later modelled as Scenario I. Section 3 introduces the other two scenarios and the use of virtual agents. Section 4 shows the validity and speed of prediction in these three scenarios for the average probability-based approach, Section 5 for the regression-based approach. Section 6 addresses the issues of cue weights, inferential statistics, and a stopping criterion that can be used in an ongoing labelling process. Finally, Section 7 concludes the paper.

## 2. Behavioural experiment

The behavioural experiment tested the roles of intonation, voice quality, and duration for the interpretation of rhetorical questions. This experiment orthogonally crossed accent type (two levels: Accent 1 vs. Accent 2)[2], voice quality (two levels: breathy vs. modal voice), and duration (two levels: long vs. short): it replicates the study by Kharaman et al. (2019) with the same number of participants and experimental lists, but without the presentation of visual stimuli. Listeners classified each 64 auditorily presented lexically ambiguous *wh*-questions as rhetorical (RQ), information-seeking questions (ISQ), or other on a button box.

### 2.1. Methods: Participants, materials and procedure

Sixteen monolingual German listeners participated in the study (12 female, 4 male, average age: 23.3 years, SD = 2.9 years). The test set consisted of 32 *wh*-questions (lexicalisations), all containing the *wh*-word *wer* 'who' (e.g., *Wer mag denn Vanille*? 'Who likes vanilla?'). The target questions were manipulated by fully crossing three prosodic factors: (a) two types of nuclear accents (Accent 1 and Accent 2) on the sentence final noun with (b) two types of voice quality (breathy and modal) on the sentence-final noun, and (c) two types of durations (long and short, i.e., 10% lengthening or 10% shortening of the whole sentence), resulting in eight combinations of cues and therefore eight test conditions.[3] Eight experimental lists were created with a pseudo-randomised order. They contained all 32 items with pitch accent type and voice quality manipulated in a *Latin-Square Design* and duration manipulated between items. Each participant was assigned to two of the lists, i.e., they heard each item twice in two of the eight conditions resulting in 64 experimental trials overall.

Prior to the experiment, participants were informed about the difference in discourse functions between ISQs (which serve to obtain information) and RQs (which serve to make a point) and were presented with a lexically unambiguous example of each illocution type (ISQ: What time is it?; RQ: Who likes paying taxes?). They were further informed that the questions presented in the experiment might be less obvious and would need to be judged based on how they sounded. To probe that the questions could indeed be interpreted as ISQ or RQ, we gave participants the option to respond "other". The participants were then seated in front of a computer screen with a button box with three response options. The buttons were labelled with the possible answers "real question", "other", and "rhetorical question" (cf. Kharaman et al., 2019). The auditory stimulus was presented over headphones and the participants had to respond as accurately and quickly as possible and classify the stimulus as "RQ", "ISQ" or "other".

### 2.2. Results and discussion

Only 4.1% of the trials were judged as "other". The participants mainly used the options "ISQ" or "RQ" as response (95.9%). The small proportion of "other" responses suggests that the prosodic realizations used in the experiment could be interpreted as either ISQ or RQ in the vast majority of cases. The "other" responses were excluded from the statistical analyses.

Fig. 2 shows the proportion of RQ-responses, i.e., the proportion of cases in which a question was judged as rhetorical in the eight test conditions. Fig. 2 suggests that there is an "additive" effect of the cues, meaning that the proportions of RQ judgements was the highest if the question was realised with Accent 1, had long duration, and breathy voice quality (93% RQ-responses for this cue combination). The proportion of RQ-responses was reduced when one of the factors changed, resulting in an outcome pattern that resembled a staircase. The role of a particular cue (in isolation and in relation to other cues) can be established via logistic regression models (cf. Schertz and Clare, 2020). In fact, the additive effect of the three cues was supported by logistic mixed-effects regression model. For this analysis, responses to "rhetorical questions" were coded as 1, responses to "real questions" as 0. The model used for the

[2] Prosodically, Accent 1 refers to a late-peak nuclear accent followed by a low boundary tone, L*+H L-% (Grice et al., 2005), while Accent 2 is an early-peak nuclear accent (H+!H*) also followed by a low boundary tone.

[3] The stimuli have originally been used in Neitsch et al. (2018) and Kharaman et al. (2019).

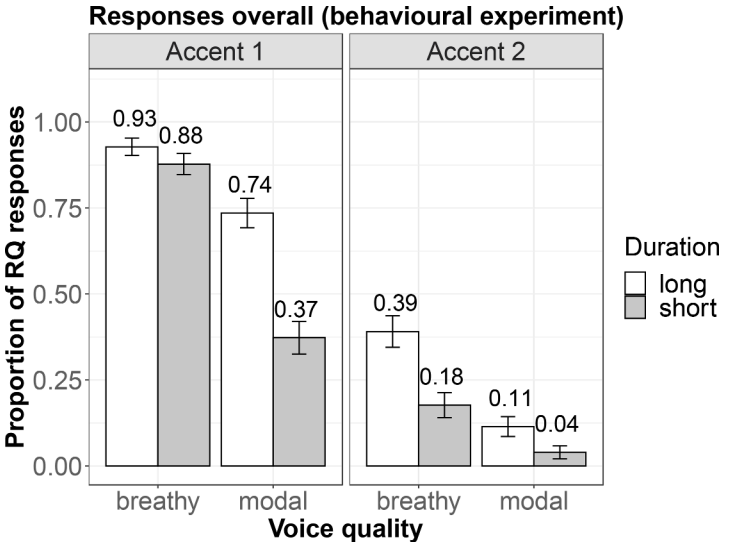



**Fig. 2.** Proportion of responses to RQs in questions with Accent 1 (panel A) and Accent 2 (panel B), split by voice quality and duration. Error bars represent +- 1 standard error of the mean.

analysis of the present data included three Factors: F1: *accent type*, F2: *voice quality*, and F3: *duration condition,* and *subjects* and *items* as crossed random factors (Baayen et al., 2008). Results showed main effects for each of the factors on the responses, see Table 1. The three factors did not interact (all two-way interactions $p > 0.2$, three-way interaction $p > 0.1$). The regression weights indicate that *accent type* ($\beta = 3.39$) has a stronger weight than *voice quality* ($\beta = -2.05$), followed by *duration condition* ($\beta = -1.28$). A post-hoc simulated power analysis (using the R-package simR, Green and MacLeod, 2016) showed a statistical power over 95% for each of the three factors.

Fig. 2 shows the proportion of RQ responses: If a question is produced with Accent 1, breathy voice quality, and long duration, it is most likely to be interpreted as RQ. In contrast, if a question is produced with Accent 2, modal voice quality, and short duration, it is most likely interpreted as ISQ.

## 3. Scenarios

In this paper, we work with three scenarios. Scenario I is the pattern derived from the behavioural experiment (Fig. 2). A scenario with three main effects and no interactions may be easy to predict because the change in one cue results in a similar decrease or increase in the probability of that cue *independently* of other cues. Scenarios II and III are entirely hypothetical, explicitly constructed for testing the AL system. In Scenario II (see Fig. 3), the weight of one cue depends on the level of another cue. That is, the weight of *duration condition* depends on the level of *voice quality* (higher proportions for Category X for stimuli with long duration when voice quality is breathy (the two left white bars in each facet) compared to when voice quality is modal (the two right white bars in each facet)).[4] Scenario III is even more complex: Here, the interpretation of a stimulus as Category X depends on three simultaneous cues (three-way-interaction): A stimulus is only interpreted as Category X when Accent 1 coincides with long duration and breathy voice quality *or* when Accent 2 coincides with shorter duration and modal voice quality. In other words, changing *duration condition* from long to short can both decrease (Accent 1, breathy voice quality) and increase (Accent 1, modal voice quality) the probability of the question to be labelled as Category X.

## 4. Active learning system with average probability-based approach

### 4.1. Methods: Active learning system, materials and procedure

Since the use of participants is time- and cost-intensive, we chose virtual agents to simulate participants' behaviour in the three scenarios.[5] This is achieved by the implementation of a binomial function. In particular, for a given cue combination of a trial we draw a random binary variable, such that the probability of occurrence of each category corresponds to the mean proportion of each

[4] We keep the factors and factor levels from Scenario I but use "Category X" and "Category Y" as response categories for all scenarios from now on to avoid confusion between hypothetical/derived scenarios and the behavioural experiment.

[5] With real participants, we would have had to implement a different experiment (with different cues or label instructions) to obtain the intended patterns. Since "other" responses were rare in the behavioural experiment, we only operated with two answer options (Category X and Y). In principle the label "other" can be included to the model by replacing the binary response model with multilevel logit or probit.

**Table 1**
Lmer output for statistical analysis of the replication experiment of Kharaman et al. (2019) study. The columns show the regression coefficient estimate (ß), the marginal effect (dF), the standard error (SD), z- and p-values. An asterisk in the last columns indicates statistical significance at 0.05 (*), 0.01 (**) and 0.005 (***).[10]

| | ß | dF | SD | z | p-value | |
|---|---|---|---|---|---|---|
| F1: Accent 1 | 3.39 | 0.54 | 0.21 | 15.43 | <0.0001 | *** |
| F2: Modal | −2.05 | −0.33 | 0.20 | −10.06 | <0.0001 | *** |
| F3: Short | −1.28 | −0.25 | 0.19 | 6.93 | <0.0001 | *** |

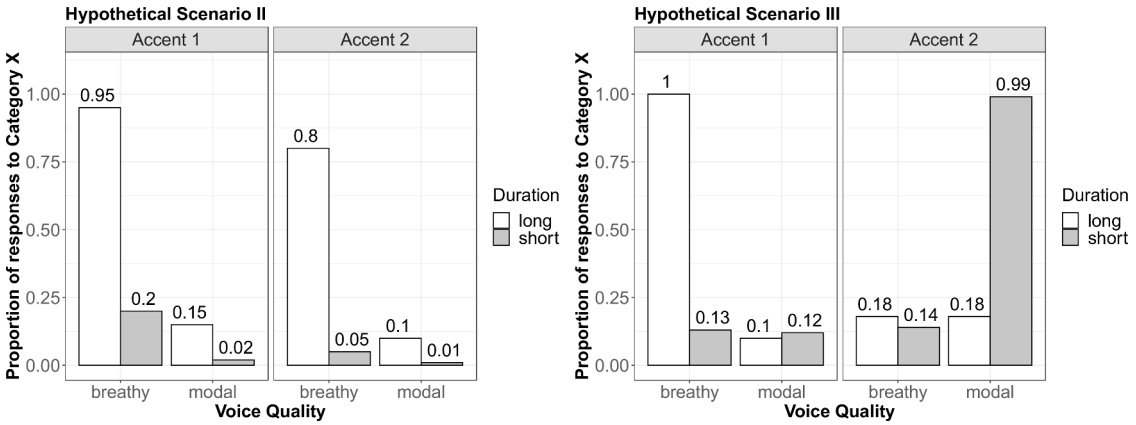



**Fig. 3.** Two different hypothetical scenarios for a 2×2×2 design. Scenario II (left panel) shows a two-way interaction between two of the three factors, and Scenario III (right panel) shows a three-way interaction between all three factors.

condition. For example, in 95 out of 100 times the virtual agent would label a cue combination {Accent 1, breathy voice quality, and long duration} as Category X (see Fig. 3, left panel).

In Section 4, we apply an Association Rule Mining algorithm (Zhang et al., 2008) to extract single cues and cue combinations that occur in at least one question instance (henceforth called *components*) and average the resulting probabilities for these components (average probability-based approach). We model a question stimulus as a set of seven components:

(a) three single cues, one for each factor (accent type; voice quality; duration condition).
(b) three combinations of two cues (accent type and voice quality; accent type and duration condition; voice quality and duration condition).
(c) one combination of three cues (accent type, voice quality, and duration condition).

An example set for the combination {Accent1, long, breathy} would be (i) {Accent1}, (ii) {long}, and (iii) {breathy} used for the single cues; (iv) {Accent1, long}, (v) {Accent1, breathy}, and (vi) {breathy, long} used for the combinations of two cues, and, finally, (vii) {Accent1, long, breathy} for the combination of three cues.

For prediction making and stimulus selection purposes, we apply a measure called *confidence,* which is commonly used in rule-mining (Agarwal and Srikant, 1994). Confidence is a conditional probability denoting the likelihood of a specific component to have a particular class label. Henceforth, we will refer to this measure using the term *probability*, see (Eq. (1)).

$$P(component,\ label) = \frac{\mathrm{P}(\text{component} \cap \text{label})}{\mathrm{P}(\text{component} \cap \text{any label})} \tag{1}$$

For instance, the probability of the component {Accent1, long} to belong to Category X class is measured as (Eq. (2))

$$\frac{\mathrm{P}(\{\text{Accent1}, \text{long}\} \text{ labelled as Category X})}{\mathrm{P}(\{\text{Accent1},\ \text{long}\} \text{ labelled as Category X or Category Y})} \tag{2}$$

Hence, for all components in the classification model, there is a probability value indicating its likelihood of belonging to Category X class. When a stimulus is labelled (as Category X or Category Y), the probability values for the seven components (i-vii), representative for the particular instance, are updated. Components of which probabilities were updated at least once are called *observed components*. The probability values of the observed components, forming the stimulus, are used to determine the probability of the stimulus to belong to the Category X class. This is done by averaging the probability values of the observed components for each stimulus after each labelling step. As foreshadowed in Section 1.1, this probability is used for two purposes:

(1) Class label prediction: We use the learned probabilities of the stimuli for prediction making. If the probability of a stimulus for Category X class is $> 0.5$, the predicted label is "Category X"; if the probability is $< 0.5$, the predicted label is "Category Y". Stimuli with a probability of 0.5 are labelled as "other". As the probability is calculated from the observed components, the model is able to predict class labels for yet unlabelled stimuli if at least one of their seven representative components has been learned/seen by the model via at least one labelling iteration. This is an attractive feature when there are many cue combinations.
(2) Stimulus selection: At the beginning of the labelling process, all stimuli are assigned 0.0 probability for all classes. At first, the model queries labels for stimuli with unique, not yet observed cue combinations. Later in the labelling process, the system queries labels for stimuli for which the classification model struggles to make predictions. That is, the system retrieves a label for a stimulus with the current average probability value that is closest to 0.5 (i.e., the most uncertain stimulus). If multiple stimuli have the same probability value, the stimulus is selected randomly from one of those.

We used 16 virtual agents to have a similar statistical power as in the behavioural experiment. Since stimulus selection was based on the AL-predictions using an uncertainty-based method, each agent received a different "experimental list" and each condition (combination of cues) could differ in the frequency of presentation. Note, the virtual agents assign a label (Category X or Y) based on the pre-defined probabilities. However, we do not allow for item or agent heterogeneity, therefore the labels from the virtual agents are generated via independent, identically distributed draws of a binomial random variable (see Fig. 1).

### 4.2. Results of the AL model's performance (average probability-based approach)

The AL system provides two kinds of outputs, the actual responses of the virtual agents (a probability for one of the response options, here Category X, averaged across the "responses" of the 16 virtual agents)[6] and the resulting predicted probability, which is updated after each response. Fig. 4 shows the gold standard, the averaged actual responses of the virtual agents after 64 trials (top panel).

The numbers below the actual response bars indicate the number of items that have been labelled by the 16 agents, summing up to 1024 responses at trial 64. The different numbers per condition are due to the uncertainty-based stimulus selection of the AL system. In Scenario I, the *uncertain* conditions (closest to an average probability of 0.5) in the centre of the eight bars (in particular bars 4 and 5, counted from the left) were presented more often than the *certain* conditions at the margins, with probabilities closer to 1 (bars 1–3) or 0 (bars 6–8). Note, however, that these numbers only provide a rough measure of what happened over the preceding trials. For instance, conditions that are *less uncertain* at trial 64 may have been *more uncertain* at earlier stages.

We now turn to the validity of the prediction and the speed with which a valid prediction is achieved. To evaluate the validity of the AL-prediction we compared the root mean squared error (RMSE) between the predicted probability of the AL system at each trial to a gold standard (the average probability of the actual responses at trial 64, see upper row of Fig. 4). The development of the RMSE between predictions and gold standard across trials is shown in Fig. 5.

Fig. 5 indicates a very stable pattern of RMSE for all three scenarios, suggesting little improvement across trials. However, it is also evident that Scenario III has considerably higher RMSE values than Scenarios II and I, a difference of more than 0.1 at trial 64 (0.29 compared to 0.15 and 0.12, respectively). Scenarios I and II are comparable in RMSE, despite the different contribution of the cues (additive in Scenario I, interacting between two cues in Scenario II). Hence, Scenarios I and II are predicted considerably better than Scenario III. The average RMSE for the average probability-based approach across all three scenarios at trial 64 was 0.19.

To determine the speed at which a first stable prediction is reached, we first calculated the proportional change in error relative to the average error in the preceding five trials. Fig. 6 visualizes this proportional change across trials. It shows the absolute (i.e., positive) value of the proportional difference in RMSE of any trial relative to the average of the preceding five trials. So, if trials 1–5 had an average RMSE of 0.17 and the sixth trial an RMSE of 0.173, then the proportional difference would be ($|0.003|/0.17 = 0.02$). When the proportional RMSE of a scenario falls below the 5%-difference-threshold for a series of five consecutive trials, the first of these trials is regarded as the stabilisation trial: trial 25 for Scenario I, trial 22 for Scenario II, and trial 11 for Scenario III. The predicted probabilities at the stabilisation trials for the respective scenario are presented in the lower panel of Fig. 4. It shows that the general pattern found at the gold standard (upper panel) is very well reproduced, but that there is a considerable regression toward the mean (probability of 0.5) for all three scenarios, but in particular for Scenario III.

### 4.3. Interim discussion of the average probability-based approach

We compared three different scenarios, a simple one with no interaction between cues (Scenario I) and successively more complicated interactions between cues (Scenarios II and III). Regarding validity, Scenario III had a higher error between predicted response probabilities and actual responses at trial 64 than Scenarios I and II (Fig. 5). Regarding the speed with which a stable pattern for Scenarios I and II was achieved, we observe a marginal change (less than 5%) from trial 25 and 22 respectively onwards (Fig. 6), suggesting that this is the earliest point where the AL reproduces the response pattern at a constant low level. Note that Scenario III

[6] Note that the labelling performed by the virtual agents (Fig. 2, Actual responses), which is based on the proportions obtained in the behavioural classification experiment, reproduces the original patterns obtained in the classification experiment quite closely. The average root-mean-squared error between actual and predicted responses was smaller than 0.03 at the end of the labelling (trial 64).

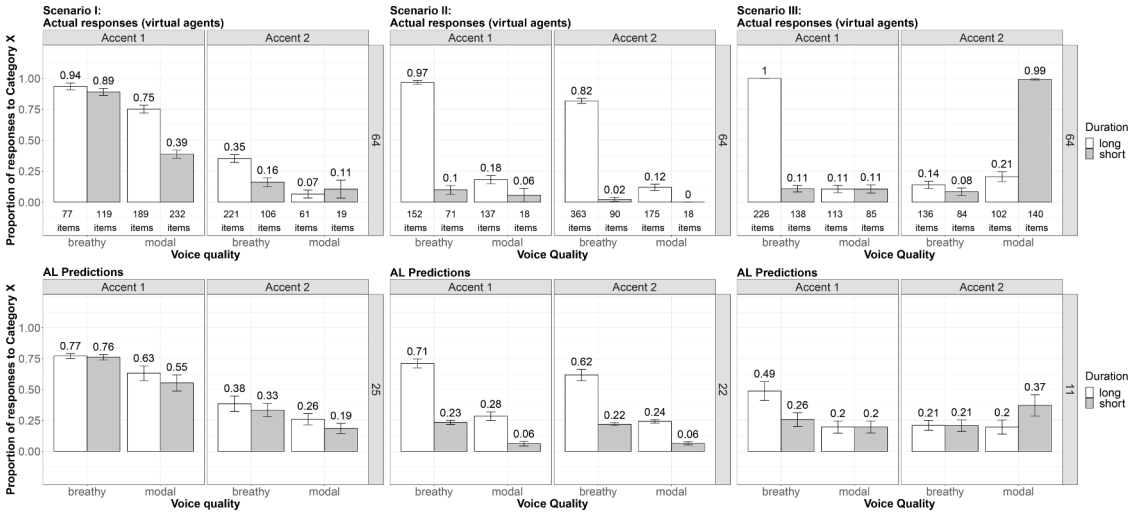



**Fig. 4.** Average probability-based AL system. Actual responses of virtual agents for trials 1–64 (gold standard, top panel) vs. AL predictions at the stabilisation trial (bottom panel) for Scenario I (trial 25, staircase) and two interactive Scenarios II (trial 22, two-way interaction) and III (trial 11, three-way interaction). *Accent type* is split in different panels (Accent 1 on left, Accent 2 on right), *voice quality* split for breathy and modal on the x-axis; and *duration* is colour-coded (long is white and short is grey). Whiskers show the standard error (SE) of the mean.

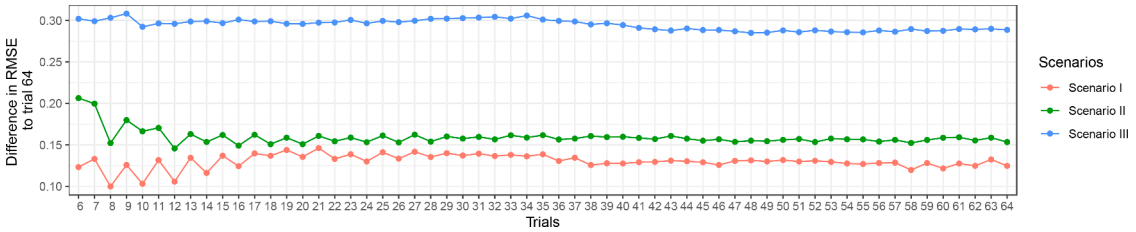



**Fig. 5.** Development of proportional differences in RMSE of the predicted response relative to the gold standard (actual responses at trial 64) and smooth across the average of the previous five trials for the three Scenarios in the average-based approach.

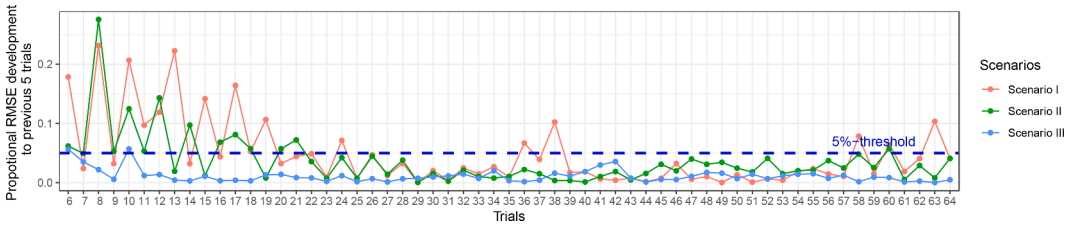



**Fig. 6.** Development of proportional differences in RMSE to the previous five trails over all 64 trials for the three Scenarios in the regression-based approach. The dashed line indicates the threshold of 5%.

becomes stable at an earlier trial (trial 11), but with a substantially higher error (see Fig. 5).

Interestingly, it was not the mere presence of an interaction between cues that made prediction harder, but the presence of an interaction between *all three cues*. In the average probability-based approach, the AL system has imposed an equal contribution of each cue and each cue combination to the probability that a given question will be classified as Category X. For the 2×2×2 design, we hence used seven different entries which contribute to the probability that a given question will be labelled as Category X (i-vii as described in Section 4.1.) and each component/cue combination thus has an equal contribution of 1/7. However, the contribution of different components/cue combinations is not necessarily equal, as there might be some cue that increases the probability of a question to be classified as Category X for a certain cue combination and decreases the probability for the scenarios with interacting cues. In the following section, we present a regression-based approach which takes potential differences in the factors' contributions into account.

## 5. Active learning system with regression-based approach

The difficulty of the average probability-based approach of Section 4 could be mitigated if the equal probability weights of each cue combination of 1/7 were adjusted, such that the AL system can differentiate the contribution of different factors for the respective category. For example, for Scenario III (right panel of Fig. 3) we would like to assign a positive weight on {long duration} only for {Accent1} and {breathy} voice quality. This can be achieved by computing the probability weights by means of a linear regression model. As will be shown below, the regression-based approach of the AL algorithm only relies on the stimuli for which labels were obtained and does not require the a-priori specification of cue weights. This is a particularly appealing feature if one does not know whether or not cues interact. Another advantage of a regression-based approach is its ability to extrapolate patterns to unseen conditions from a fewer number of initial labels. Given these advantages, a regression-based approach was implemented in order to optimise the AL's prediction.

### 5.1. Methods: Linear regression-based approach

A linear regression model is a way to adjust the weights ($w$) of different components/cue combinations to the probability of a stimulus to be classified as Category X. It still assumes a linear dependence but allows the contributions of each component to differ from a fixed weight of 1/7:

$$P(F_1,\ F_2,\ F_3) = w_0 + w_1F_1 +\ w_2F_2 +\ w_3F_3 +\ w_4F_1F_2 +\ w_5F_2F_3 +\ w_6F_1F_3 +\ w_7F_1F_2F_3 \tag{3}$$

where $F_1,\ F_2,\ F_3$ ($F =$ factor) denote the values of Factor 1, 2, and 3 respectively (these are encoded as binary $\{0, 1\}$ values and the corresponding contributions of the components/cue combinations $\{w_0,\ \ldots,\ w_7\}$, which are now flexible and may differ from 1/7). The weights are inferred from the labelled data with the ordinary least squares procedure (OLS, e.g., Wooldrigde, 2016). The idea of OLS is to minimise the squared distance between the estimated probability of a question to be a Category X, $P(F_1,\ F_2,\ F_3)$ and the actual label received from a participant, by changing the weights ($w$). Once the weights are calibrated based on all available responses, the probability of a question with any cue combination can be computed accordingly by inserting the values of the Factors 1, 2 and 3 to Eq. (3) (see Section 6.2 for a worked-out example). This implies that the regression adjustment of the AL system makes the algorithm less dependent on pre-set (initial) probability values. Moreover, this generalised probability definition is flexible enough to allow for cues to have more than two values. The estimated weights $w$ can take on any value, including negative ones. This allows us to overcome the problem of Scenario III in which cue effects point in opposite directions, e.g., if we encode $(F_1,\ F_2,\ F_3) = (1, 1, 1)$ for a question with "Accent 2″, "modal voice quality" and "short duration", the model would assign a positive value for $w_7$ and negative values for $w_1$, $w_2$ and $w_3$. This assures that the contribution of "short duration" is positive only for this specific cue combination $(F_1,\ F_2,\ F_3) = (1, 1, 1)$ and is negative for the other conditions. Such model flexibility comes with the price of being unstable for smaller numbers of trials. It has been argued in the literature that the best option is to truncate predicted probabilities that lie outside the [0,1] interval with zero and one respectively (Lee et al. (2011) and references therein).[7]

The AL approach is equivalent to the approach used for the average-probability model (described in Section 4): (1) We use the learned probabilities of the stimuli from the regression model for prediction making. If the probability of a stimulus for Category X class is $> 0.5$, the predicted label is Category X; if the probability is $< 0.5$, the predicted label is Category Y. Otherwise, the data are excluded. (2) At the beginning of the labelling process, all stimuli are assigned 0.0 probability for all classes. At first, the model queries labels for stimuli with unique, not yet observed cue combinations. Next, it retrieves labels for the most uncertain stimuli (i.e., where the probability for the prediction is close to 0.5).

### 5.2. Results of the active learning model's performance (regression-based approach)

Fig. 7 gives an overview of the actual responses given by the virtual agents at trial 64 (gold standard, top panel) and on the bottom panel the predictions made by the AL based on the regression-based approach at the respective stabilisation trial for each scenario (Scenario I: trial 26, Scenario II: trial 14 and Scenario III: trial 20, see Fig. 9).

For the analysis of the validity of the regression-based AL system, the procedure was identical to that described in Section 4.1. Fig. 8 shows the development of the error relative to the gold standard (actual response probability at trial 64) across trials.

Unlike Fig. 5 for the avarage probability-based approach, Fig. 8 shows a decrease in RMSE across trials, which varies in its steepness. A large reduction in error is already observed in the first 20 trials (Scenario I: 6–19 from 0.18 to 0.12, Scenario II: 6–8 from 0.20 to 0.14, Scenario III: 6–16 from 0.28 to 0.2). After that, a phase of stabilisation is visible for all scenarios for the next 20 trials at least. While Scenario I continues on this plateau with an average error of 0.16, the error for Scenario III drops considerably between trials 40 and 42 (from 0.20 to 0.15) and the same is true for Scenario II between trials 52 and 53 (from 0.12 to 0.08). The average RMSE

[7] As an alternative solution, one could avoid the hard truncation by implementing a nonlinear logistic regression model. A logistic regression is similar to linear regression, as it also controls the probability change by varying the weights $w$, however, for the binary factors, or the 2x2x2 design we have considered, the results of truncated linear regression and the logistic regression are numerically identical. The logistic regression is extremely useful for multilevel cues, but for our particular design, we leave the linear form of Eq. 3, which is more straight-forward to interpret than logistic regression.

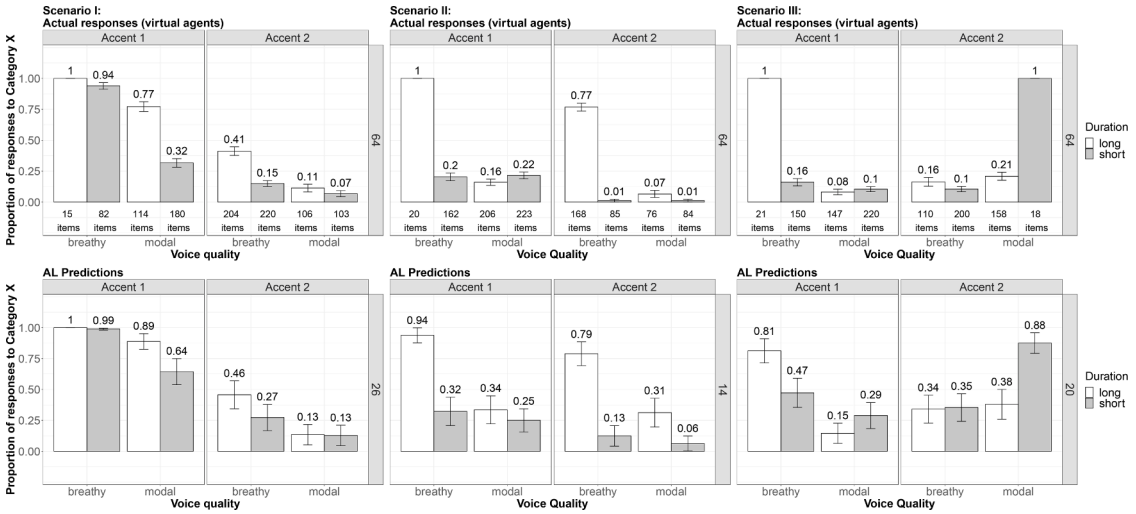



**Fig. 7.** Regression-based AL system. Actual responses of virtual agents for trials 1–64 (gold standard, top panel) vs. AL predictions at the stabilisation trial (bottom panel) for Scenario I (trial 26, staircase) and two interactive Scenarios II (trial 14, two-way interaction) and III (trial 20, three-way interaction). *Accent type* is split in different panels (Accent 1 on left, Accent 2 on right), *voice quality* split for breathy and modal on the x-axis; and *duration* is colour-coded (long is white and short is grey). Whiskers show the standard error (SE) of the mean.

was smaller in the regression-based approach than the average-probability approach in Fig. 5 for all scenarios (0.15 vs. 0.19).

Fig. 9 shows the proportional change in RMSE compared to the average of the previous 5 trials. The first trial in a row of five consecutive trials with errors less than 5% is trial 26 for Scenario I, trial 14 for Scenario II, and trial 20 for Scenario III.

Finally, we compare the validity of the prediction across the two approaches, probability-based and regression-based directly: To this end, we divided the RMSE values of the regression-based approach by the RMSE values of the average probability-based approach (comparison of Figs. 7 and 4) at the stabilisation trial for each scenario (see Fig. 9). The RMSE in the regression-based approach is lower (i.e., better) or equally good for all three scenarios at the trial where it meets the speed criterion in the regression-based approach compared to the average probability-based approach: The RMSE of the regression-based approach for Scenario I at trail 26 is only marginally lower, the RMSE for Scenario II at trial 14 is 18% lower and the RMSE for Scenario III at trail 20 is 33% lower than that of the average probability-based approach respectively.

### 5.3. *Interim discussion of the regression-based approach*

The analysis of the validity suggests that there is improvement in prediction error especially in the first third of the experiment, followed by a phase of stability and another improvement for Scenarios II and III. The improvement of the regression-based approach over the average probability-based approach is especially noticeable for Scenario III, which was difficult to predict for the average probability-based AL system and whose error did not fall considerably below 0.3 (see Fig. 5). The RMSE values are on avarage lower in the regression-based approach from the start to the end of the experiment, in particluar for Scenario III. This was also confirmed by calculating an RMSE ratio at a certain trial between both approaches to evaluate the goodness of fit of different models the RMSE values by comparing them.

Overall, the regression-based approach shows a reduction in RMSE throughout the experiment, with larger changes in the first third of the experiment, and, for Scenarios II and III again in the last third.

The speed analysis revealed a slightly later stabilisation of the regression-based approach compared to the average-probability approach, but given the higher validity, this is affordable. Taking a threshold of a relative change in error relative to the preceding five trials of 0.05, the model is locally stable after trial 26 the latest. Interestingly, Scenario I, without interaction between cues, took

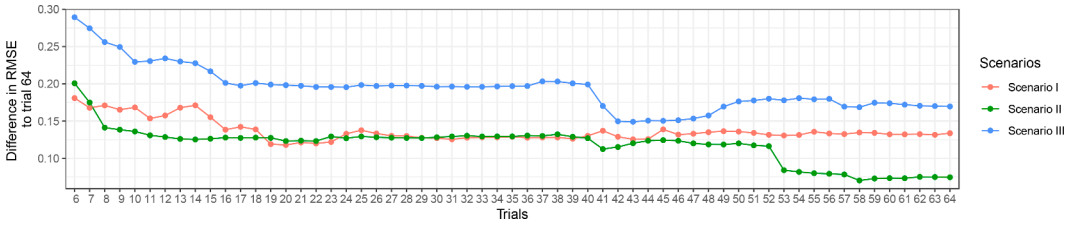



**Fig. 8.** Development of proportional differences in RMSE of the predicted response relative to the gold standard (actual responses at trial 64) and smooth across the average of the previous five trials for the three scenarios in the regression-based approach.

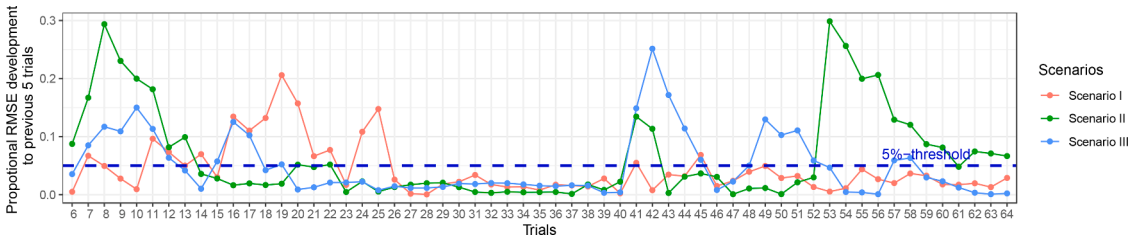



**Fig. 9.** Development of proportional differences in RMSE to previous five trails over all 64 trials for the three Scenarios in the regression-based approach. The dashed line indicates the threshold of 5%.

longest to stabilise and did reach highest accuracy (residual error of 0.16 at trial 64, compared to 0.07 for Scenario II). Scenario II, in turn, stabilises from trial 14 onwards, i.e., after slightly less than a ¼ of the planned labelling process and shows only a marginal gain until trial 50. Scenario III paints a similar picture: The proportional difference trajectory stabilises at trial 20 and shows only marginal changes until trial 40, see Fig. 9.

Taken together, the regression-based approach led to lower errors than the average-probability approach at the end of the experiment, averaged over all scenarios. Only Scenario I (with main effects only) was predicted slightly better in the average-probability based approach (RMSE of 0.12 compared to 0.13). Given that the experimenter does not know a-priori which pattern to expect, the regression-based approach seems to be a sensible choice because it maximizes accuracy.

## 6. Applying active learning systems for psycholinguistic research

The evaluation showed that an AL system with a regression-based approach is able to predict the actual responses at the end of the experiment with good validity (average error of 0.13 at the end of the experiment). For actual application in psycholinguistic research, we discuss the use of a stopping criterion (6.1) and the extraction of cue weights and inferential statistics (6.2).

### 6.1. Stopping criterion

The proportional RMSE analysis (speed analysis in Sections 4.2 and 5.2) could serve as a stopping criterion for the labelling process and end the experiment for a participant once the average deviance of the responses has reached the 5%-threshold for a series of five consecutive trials. Given the way we operationalized the stopping criterion, i.e., the stabilisation trial, so far, the analysis is only feasible a-posteriori, i.e., after the experiment has been completed. A real-time stopping criterion in an on-going experiment can be defined in terms of the stabilisation of the predicted probabilities of the AL system (rather than the RMSE which has the nature of a post-hoc measure in our case). A conservative threshold could be a 5%-divergence as used before, but a more sensitive 1%-threshold may be sensible, depending on the task. We show the use of a stopping criterion with the 5%-threshold for the three scenarios of the regression-based approach in Fig. 10. The stopping criterion would be reached at trial 21 for Scenario I[8], at trial 14 for Scenario II and at trial 18 for Scenario III. This means the stopping criterion would be reached even earlier than the stabilisation of the RMSE for Scenarios I (RMSE-based stabilisation: trial 26) and III (RMSE-based stabilisation: trial 20) and at the same point for Scenario II (RMSE-based stabilisation: trial 14). Overall, the speed for both criteria is comparable.

As Fig. 10 shows, the proportional predicted AL probabilities do not necessarily decrease with the number of trials. Participant heterogeneity might lead to the fact that a very different style of labelling occurs at the later stages of the experiment. We address this issue by considering the predicted AL probabilities' improvements averaged over 5 trials. This also means that changes later in the experiment are rather minimal in absolute terms. However, if one expects high participant heterogeneity it is advisable to increase the number of trials to compute the improvement of the average predicted AL probabilities. Another factor for the higher predicted AL probability values towards the end of the experiment could be the uncertainty-based approach we chose for the stimulus selection. Participants are confronted with the most uncertain cue combinations, i.e., presumably most difficult stimuli to categorize, towards the end of the experiment which could lead to a less stable pattern. We will discuss solutions to this in Section 7.

### 6.2. Cue weights and inferential statistics

In an actual behavioural experiment, the importance of cues and cue combinations can be established with a logistic mixed-effects regression model (in terms of main effects, interactions and coefficients), cf. Schertz and Clare (2020) and the analysis in Section 2.2. The model in Section 2.2. included three factors (*accent type, voice quality*, and *duration condition*) as fixed effects and *participants* and *items* as crossed random factors (Baayen et al., 2008). The inclusion of random effects allows to generalize the contribution of the fixed effects beyond the particular participants and items of the experiment. The situation is different for the AL model parameters in the

[8] If the average value of the preceding 5 trial equals 0, NA was inserted. This was the case for the cue combination {Accent 2, breathy, short} between trials 46-51 of Scenario I. This means that the values did not contribute to the average plotted in Fig. 10.

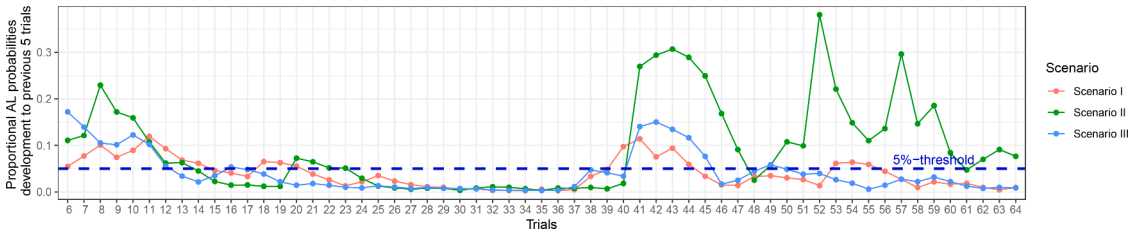



**Fig. 10.** Proportional development of the predicted AL probabilities compared to the average of the previous 5 trials of Scenarios I-III, regression-based approach. The dashed line indicates the threshold of 5%.

simulations. First, the virtual agent's decision at each trial was an independent draw from a binomial function (and is hence different from idiosyncratic behaviour of participants that needs to be modelled with random effects). The same is true for the different lexicalisations, which are ignored by the agents (but may affect participant's reactions in a real experiment). Adding random effects in our case makes the model unnecessarily complex and invokes estimation noise to the relevant parameters. Second, mixed effects models are needed when coefficients of the models are expected to depend on a certain group of observations, e.g., items or participants. With AL, a given participant will be presented a *certain* cue combination only a few times and an *uncertain* cue combination might be presented for several items. Therefore, the standard generalized linear mixed effects models, which assume the same linear structure of random effects for all observations, are no longer applicable.

As an example, for conducting inference on the estimated cue weights, consider Scenario I in the regression-based approach. Note that we use a generalized linear model (glm) in this example, rather than a mixed-effects model (glmer) because the judgements from the virtual agents came from independent draws of the binomial distribution and were hence not heteroscedastic. The data used in the analysis below corresponds to 21 trials[9], i.e., the trial where the stopping criterion would first apply (see Section 6.1, Fig. 10). Table 2 provides the output of the most parsimonious glm that includes only significant effects (using a glmer leads to the same output). The first column reports the estimates of the regression coefficients, whereas the marginal effects (dF) of each individual cue are reported in the second column. The relative importance of cues and cue combinations/components can be extracted from the marginal effects, i.e., be read off from the second column: changing duration from *long* to *short* would decrease the probability of the stimulus labelled as Category X by 31%. If we had interactions between cues, to determine the weight of a single cue, e.g., *Accent* (F1), we would have to fix the other cues to a neutral level, e.g.$(F_2 = \ F_3 = 0.5)$, and add up the products (ß(F1) + 0.5 * ß(F1 * F2) + 0.5 * ß(F1 * F3) + 0.5 * ß(F1 * F2 * F3)).

The statistical result after trial 21 was hence identical to the behavioural experiment in Section 2: all three factors *accent, duration condition*, and *voice quality* have an effect on the probability of a stimulus being judged as Category X (all $p < 0.05$). The marginal effects were similar; the AL model slightly overestimated the effect of *accent* (0.58 vs. 0.54) and *duration condition* (0.31 vs. 0.25).

The statistical analysis presented here is for the analysis of the virtual agents' data (in which there is no item or participant heterogeneity). In an application with human labellers, the glm would have to be changed to a mixed effects model (glmer) in order to account for random effects caused by participants and items. Other than that, the analysis would be similar.

## 7. Discussion, conclusion, and outlook

We examined whether AL systems as a methodological tool can be successfully employed in psycholinguistics, which was tested by the example of a cue weighting study, and – if so – which approach is better suited to predict the actual responses, an average probability-based approach or a regression-based approach. These two approaches have different advantages. Since it is unclear a-priori whether cues will interact (trading relation) or not, we tested three hypothetical scenarios (Scenario I with main effects of each cue and no interactions, Scenarios II and III with trading relations, i.e., interactions between two and three cues respectively). All of these scenarios may be realistic results of psycholinguistic experiments. We showed that AL systems with a regression-based approach were overall better-suited to predict the outcome of the experiment in all three scenarios. A good approximation (RMSE $< 0.20$) was already reached after 1/3 of the original experiment (trial 21) for all scenarios. The average error across all three scenarios dropped to 0.12 compared to 0.19 in the average probability-base approach. The early stabilisation of errors suggests a quick estimation of the probabilities. This stabilisation could be used as a stopping criterion in a larger experiment, if necessary (Section 6.1). Needless to say, the longer the labelling process, the smaller the RMSE values. This was especially true for the interaction scenarios. Scenario II reaches the global minimum at trial 58 (RMSE 0.07), i.e., at the end of the labelling process. In sum, the AL system allows us to reduce the number of trials in an experiment, which is attractive in the case of very complex study designs (with many different cues, cue levels and lexicalisations).

From a computational perspective, the present study is informative as to how different approaches (average probability-based vs.

[9] An output like Table 2 can principally be retrieved at any point during the labelling process. Yet, only those cues that change within the range of trials can be included. For, instance, if the factor *duration condition* was *long* for all 21 trials, then the effects of the cue *duration* (main effect and interactions) cannot be calculated.

[10] dF stands for the derivative of the logistic cumulative distribution function and represents the marginal effect of a cue.

**Table 2**
Marginal effects for AL predictions after 21 trials, Scenario I, regression-based approach, output of the logistic regression model. The columns show the regression coefficient estimate (ß), the marginal effect (dF), the standard error, z- and p-values. An asterisk in the last columns indicates statistical significance at 0.05 (*), 0.01 (**) and 0.005 (***).

| | ß | dF | Std. Err. | Z | p-value | |
|---|---|---|---|---|---|---|
| F 1: Accent 1 | 4.33 | 0.58 | 0.49 | 8.90 | <0.0001 | *** |
| F 2: Modal | −2.49 | −0.31 | 0.46 | −5.44 | <0.0001 | *** |
| F 3: Short | −1.86 | −0.31 | 0.35 | −5.33 | <0.0001 | *** |

regression-based) employed in AL affect the predictions. The comparison between the average probability-based approach and the regression-based approach showed a better performance of the regression-based approach overall, but there were differences depending on the complexity of the cue interactions. In the simplest case with only main effects, Scenario I, the RMSE at trial 64 was nearly identical in the two approaches. In the more complex Scenarios II and III, RMSEs were considerably lower in the regression-based approach compared to the average probability-based approach (0.07 vs. 0.15, 0.17 vs. 0.29, respectively). In sum, while the average probability-based approach of cue contributions represented a necessary first step to test the usefulness of AL in psycholinguistic cue weighting research, it turned out to be too simplistic when there were interactions between cues. Another difficulty for the average probability-based approach may have been classifications close to 0 or 1 (Jalalzai et al., 2018). For a scenario without interactions, the choice of approach is incidental as the comparison between the RMSE values revealed no difference. In the case of interactions, the regression-based approach was more flexible. The proposed AL system with a regression-based approach is promising also for larger multi-level cue experiments. By design, the AL systems recalculate the predicted classification probabilities for all questions with a single participant label. Such increase in effective sample size allows for an implementation of logistic regression and support vector machines to further improve on classification accuracy (Friedman et al., 2001). In sum, it is important to run the AL system with the optimal internal approach; for the potential outcome scenarios tested here, the regression-based approach was clearly superior.

If one uses AL for more complex designs, the specifics of AL-internal approaches have implications for stimulus selection. Currently, the stimulus selection relies on uncertainty-based sampling. In larger-scale experiments, uncertainty-based sampling might introduce negative artefacts (e.g., imbalanced datasets) since it is not guaranteed that the system would select stimuli equally from both/all response classes. To avoid these artefacts, it is common to combine multiple stimulus selection strategies within a single labelling session. As previous AL research suggests (Bernard et al., 2018), we might apply data-based strategies to cover the data-space in the early stage of the labelling process. When a sufficient number of stimuli was labelled in order to create an initial classification model, we might apply certainty-based strategies to query labels for uncertain stimuli, that way increasing the model's confidence.

Our study leaves some questions open, mostly methodological in nature. First, one of the reasons of an increased speed in predicting the outcome of the experiment lies in the selection of conditions whose previous labelling suggests uncertainty. These are presented to the participant more often than conditions whose labelling suggests more certainty. In settings with human participants instead of virtual agents, this may lead to entrainment or to an enhanced sensitivity to contrasts that may go unnoticed when *certain* conditions are included. Also, participants may respond strategically (e.g., avoidance of pressing the same button more than 4 to 5 times in a row). There are some studies suggesting "adaptive" behaviour, either based on entrainment (e.g., Lewandowski and Nygaard, 2018; Pardo, 2006) or order effects (e.g., Lucas, 1992; Moore, 1999). It may also be the case that human participants will have to get used to the speaker's characteristics (general voice setting, speech rate, pitch range). Our underlying assumption is that actual human participants would follow a mental grammar when doing this task, which also binds them to a certain pattern – similar to the behaviour of virtual agents. This assumption seems warranted. A replication of that study in Section 2.2 with four speakers (instead of only one speaker) showed again similar main effects and cue weightings as in the current study (Geib and Braun, 2022). Another factor, which does not affect the virtual agents but might affect human participants, are the lexicalisations, as potential (individual) interpretation could influence the participant's choice apart from the pure linguistic means. In the present study, the uncertainty-based choice of the upcoming stimulus is only based on the components mentioned in Section 4.1 but does not include *lexicalisations* as factor. The information on the lexicalisations could be included in the stimulus selection procedure. From a different perspective, one may also take advantage of individual differences: One could let the system learn multiple predictive models in a single experiment, which are tailored to groups of participants that share a sensitivity to one or the other cue. For instance, one might first model the participant's profile (i.e., cues that the participant seems to pay attention to) and use an AL model that has been trained on data received from participants with a similar profile. This would increase the stability of the AL model used for predictions and stimuli selection, which is the critical element that impacts the efficacy of the whole approach. Furthermore, it would allow us to check for individual differences, by relating perceptual profiles to participants' metadata.

Second, a further open question our study cannot answer is whether the AL systems perform equally well if there are designs with even more factors or with factors that have more than two levels. Theoretically, a logistic regression-based approach could serve as a solution for designs with multilevel factors but we did not test such a design and approach in this study. However, further testing with a $2\times2\times2\times2$ study design indicates that the regression-based approach is promising even with an additional factor (Einfeldt et al., 2021).

Regarding data analysis, classical experiments rely on inferential statistics in which the effects of the different factors, potential interactions (and effect sizes) are compared, so that cue weights can be determined. The analysis we presented here is in the form of a model that assigns weights to the individual factors and their interactions. These weights can be directly compared and corroborated, as shown above, by inferential statistics (Section 6.2). In a more general setting with multilevel cues and heterogeneous participants

modern statistical methods allow to conduct valid inference, providing more insights on the learning patterns and cue weights for participants with different profiles. For instance, an AL framework allows to infer the participants' learning pattern first and build a statistical model for cue weights for each learning profile separately, which allows us to model an individual's learning pattern in a more realistic way.

From a more philosophical perspective, one may wonder how close the computational AL approach is to human learning, both in a second language (L2) and the native (L1) one. There is anecdotal evidence that learners query phonetic/phonological information on word forms they are uncertain about (e.g., an L2 learner saying [gat:o] in Italian with a half-long [t] and waiting for confirmation whether this classifies as a long consonant (geminate) or short consonant (singleton); a child pointing at a cat, saying "cat" and waiting for confirmation). In future research, it may be interesting to test whether learners who try to get feedback on uncertain cases (rather than any arbitrary case) learn faster. This could be done with the uncertainty-based stimulus selection we used in this experiment. Also, it may be interesting to test how quickly cue trading patterns can be acquired in real-life learning scenarios. One could provide participants with a certain pattern of results and see how well they can learn the underlying structure. In the Italian example, one could present learners with different words that have different vowel and consonant durations as well as speaking rates and provide the category (*singleton* or *geminate*). For analytic participants qualitative post-experiment interviews may even grant access to perceived cue weights and hence rudimentary information on the interface between phonetics and phonology (e.g., "Whenever the vowel was short, I pressed geminate, unless the person spoke really slowly, then I did so only when the consonant seemed long"). It is very likely that computational approaches are better than humans, in particular with more complex patterns.

To conclude, while AL has already been shown to be useful for different natural language processing tasks such as named-entity recognition (Shen et al., 2004) or semantic parsing (Thompson et al., 1999), we show that AL can be applied also in other linguistic subfields (e.g., in psycholinguistics experiments at the prosody-pragmatics interface) by the example of a cue weighting study. AL is particularly suited to study cue weighting with a large number of orthogonally varied cues in different lexicalisations (resulting in a large number of conditions). This cannot be easily done with standard psycholinguistic experiments, in which each participant reacts to a number of prespecified items (lexicalisations) for each condition, because this would result in too many experimental trials for each participant. At the same time, such complex designs are necessary to fully understand the interplay between various different cues for the interpretation of linguistic and paralinguistic contrasts.

The code for the experiment can be found under https://github.com/AL-perception-experiment/simulation.

## Funding

The work presented here was funded by the DFG as part of research unit 'Questions at the Interfaces' (FOR 2111, project P6 and P8), grant numbers [BR 3428/4–1/2] (Bettina Braun), [DE 876/3–1/2] (Nicole Dehé) and [KE 740/17–2] (Daniel Keim).

## CRediT authorship contribution statement

**Marieke Einfeldt:** Methodology, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Rita Sevastjanova:** Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Katharina Zahner-Ritter:** Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Ekaterina Kazak:** Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Bettina Braun:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Visualization, Funding acquisition, Writing – original draft, Writing – review & editing.

## Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Marieke Einfeldt reports financial support was provided by German Research Foundation. Rita Sevastjanova reports financial support was provided by German Research Foundation. Katharina Zahner-Ritter reports financial support was provided by German Research Foundation.

## Data availability

Data will be made available on request.

## Acknowledgments

We are grateful to the audience of satellite workshop "*Cue weighting: Thinking outside the box*" at LabPhon 2020 (Vancouver, virtual conference) for the discussion of parts of the results. We also thank the members of the Research Unit 'Questions at the Interfaces' (FOR 2111) for feedback. We also thank Mennatallah El-Assady for feedback on the Active Learning system, and Nicole Dehé for comments to an earlier version of this manuscript.

## References


Agarwal, R., Srikant, S., 1994. Fast algorithms for mining association rules. In: Proceedings of the 20th Very Large Database (VLDB) Conference.

Andreeva, B., Barry, W.J., Steiner, I., 2007. Producing phrasal prominence in German. In: Proceedings of the 16th International Congress on Phonetic Sciences (ICPhS). Saarbrücken, Germany, pp. 1209–1212.

Andreeva, B., Barry, W.J., Wolska, M., 2012. Language differences in the perceptual weight of prominence-lending properties. In: Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech). Portland, OR, USA, pp. 2426–2429.

Angluni, D., 1988. Queries and concept learning. Mach Learn 2 (4), 319–342.

Baayen, R.H., Davidson, D.J., Bates, D.M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. 59 (4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Barry, W.J., Andreeva, B., 2011. Is it important for communication which parameters signal accentuation?. In: Proceedings of the 17th International Congress on Phonetic Sciences (ICPhS). Hong Kong, China, pp. 288–291.

Baumann, S., 2014. The importance of tonal cues for untrained listeners in judging prominence. In: Proceedings of the 10th International Seminar on Speech Production (ISSP). Cologne, Germany, pp. 21–24.

Baumann, S., Röhr, C., 2015. The perceptual prominence of pitch accent types in German. In: Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS XVIII). Glasgow, UK, p. 2981.

Baumann, S., Winter, B., 2018. What makes a word prominent? Predicting untrained German listeners' perceptual judgments. J. Phon. 70, 20–38. https://doi.org/10.1016/j.wocn.2018.05.004.

Beach, C.M., 1991. The interpretation of prosodic patterns at points of syntactic structure ambiguity: evidence for cue trading relations. J. Mem. Lang. 30 (6), 644–663. https://doi.org/10.1016/0749-596X(91)90030-N.

Bernard, J., Zeppelzauer, M., Lehmann, M., Müller, M., Sedlmair, M., 2018. Towards user-centered active learning algorithms. Comput. Graph. Forum 37 (3), 121–132. https://doi.org/10.1111/cgf.13406.

Braun, A., Schmiedel, A., Winter-Froemel, E., Thaler, V., 2018. The phonetics of ambiguity: a study on verbal irony. Cultures and Traditions of Wordplay and Wordplay Research. De Gruyter, Boston, Berlin, pp. 111–136.

Braun, B., Dehé, N., Einfeldt, M., James, A., Kazak, E., Sevastjanova, R., Wochner, D., Zahner-Ritter, K., 2022. What makes a question rhetorical? Evidence from a multiple-cue perception experiment. In: Proceedings of the Presentation at Annual Conference of the Centre for Excellence in Estonian Studies. Tartu.

Cangemi, F., D'Imperio, M., 2013. Tempo and the perception of sentence modality. Lab. Phonol. 4 (1), 191–219. https://doi.org/10.1515/lp-2013-0008.

Cheang, H.S., Pell, M.D., 2008. The sound of sarcasm. Speech Commun. 50 (5), 366–381. https://doi.org/10.1016/j.specom.2007.11.003.

Chrabaszcz, A., Winn, M., Lin, C.Y., Idsardi, W.J., 2014. Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. J. Speech Lang. Hear. Res. 57 (4), 1468. https://doi.org/10.1044/2014_JSLHR-L-13-0279.

Cole, J., Mo, Y., Hasegawa-Johnson, M., 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. Lab. Phonol. 1 (2), 425–452. https://doi.org/10.1515/labphon.2010.022.

Cole, J., Shattuck-Hufnagel, S., 2016. New methods for prosodic transcription: capturing variability as a source of information. Lab. Phonol. J. Assoc. Lab. Phonol. 7 (1), 1–29. https://doi.org/10.5334/labphon.29.

Cutler, A., 1977. The context-dependence of "intonational meanings". In: Proceedings of the Papers from the 13th Regional Meeting of the Chicago Linguistic Society, pp. 104–115.

Einfeldt, M., Sevastjanova, R., Zahner-Ritter, K., Kazak, E., Braun, B., 2021. Reliable estimates of interpretable cue effects with active learning in psycholinguistic research. In: Proceedings of the 22th Annual Conference of the International Speech Communication Association (Interspeech), Brno, pp. 1743–1747.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning, 1. Springer series in statistics.

Frost, D., 2011. Stress and cues to relative prominence in English and French: a perceptual study. J. Int. Phon. Assoc. 41 (1), 67–84. https://doi.org/10.1017/S0025100310000253.

Fry, D.B., 1958. Experiments in the perception of stress. Lang. Speech 1 (2), 126–152.

Geib, L., Braun, B., 2022. Influence of speaker characteristics on the interpretation of rhetorical questions. In: Proceedings of the Presentation at Phonetik und Phonologie im deutschsprachigen Raum. Bielefeld.

Genzel, S., Kügler, F., 2020. Production and perception of question prosody in Akan. J Int Phon Assoc 50 (1), 61–92. https://doi.org/10.1017/S0025100318000191.

Gollrad, A., Sommerfeld, E., Kügler, F., 2010. Prosodic cue weighting in disambiguation: case ambiguity in German. In: Proceedings of the 5th International Conference on Speech Prosody.

Gordon, M., Roettger, T., 2017. Acoustic correlates of word stress: a cross-linguistic survey. Linguist. Vanguard 3 (1), 20170007.

Green, P., MacLeod, C.J., 2016. SIMR: an R package for power analysis of generalised linear mixed models by simulation. Methods Ecol. Evol. 7 (4), 493–498. https://doi.org/10.1111/2041-210X.12504.

Grice, M., Baumann, S., Benzmüller, R., Sun-Ah, J., 2005. German intonation in autosegmental-metrical phonology. Prosodic Typology. The Phonology of Intonation and Phrasing. Oxford University Press, Oxford, pp. 55–83.

Jalalzai, H., Clémençon, S., Sabourin, A., 2018. On binary classification in extreme regions. Adv. Neural. Inf. Process Syst. 31, 3092–3100.

Japkowicz, N., Shah, M., El Naqa, I., Ruijiang, L., Murphy, M.J., 2015. Performance evaluation in machine learning. Machine Learning in Radiation Oncology. Theory and Applications. Springer Cham, Heidelberg, New York, Dordrecht, London, pp. 41–56.

Kaland, C., 2020. Offline and online processing of acoustic cues to word stress in Papuan Malay. J. Acoust. Soc. Am. 147 (2), 731–747. https://doi.org/10.1121/10.0000578.

Kharaman, M., Xu, M., Eulitz, C., Braun, B., 2019. The processing of prosodic cues to rhetorical question interpretation: psycholinguistic and neurolinguistics evidence. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech). Graz, Austria, pp. 1218–1222.

Kohler, K., 1991. Terminal intonation patterns in single-accent utterances of German: phonetics, phonology and semantics. Arbeitsberichte Inst. Phon. Digit. Sprachverarbeitung Univ. Kiel AIPUK 25, 115–185.

Kohler, K., 2008. The perception of prominence patterns. Phonetica 65, 257–269.

Kohler, K., 2012. The perception of lexical stress in German: effects of segmental duration and vowel quality in different prosodic patterns. Phonetica 69, 68–93. https://doi.org/10.1159/000342126.

Ladd, D.R., 2008. Intonational Phonology. Cambridge University Press, Cambridge.

Lee, B., Lessler, J., Stuart, E., 2011. Weight trimming and propensity score weighting. PLOS One 6, e18174. https://doi.org/10.1371/journal.pone.0018174.

Lehiste, I., Olive, J.P., Streeter, L.A., 1976. Role of duration in disambiguating syntactically ambiguous sentences. J. Acoust. Soc. Am. 60, 1199–1202.

Lewandowski, E.M., Nygaard, L.C., 2018. Vocal alignment to native and non-native speakers of English. J. Acoust. Soc. Am. 144 (2), 620. https://go.exlibris.link/rzBpCFw7.

Lucas, C.P., 1992. The order effect: reflections on the validity of multiple test presentations. Psychol. Med. 22 (1), 197–202. https://doi.org/10.1017/S0033291700032852.

McCallum, A., Nigam, K., 1998. Employing EM and poolbbased active learning for text classification. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 350–358.

Mo, Y., Cole, J., 2010. Perception of prosodic boundaries in spontaneous speech with and without silent pauses. J. Acoust. Soc. Am. 127 (3), 1956. https://doi.org/10.1121/1.3384972.

Moore, D.A., 1999. Order effects in preference judgments: evidence for context dependence in the generation of preferences. Organ. Behav. Hum. Decis. Process. 78 (2), 146–165. https://doi.org/10.1006/obhd.1999.2828.

Nauke, A., Braun, A., 2011. The production and perception of irony in short context-free utterances. In: Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS), pp. 1450–1453.

Neitsch, J., Braun, B., Dehé, N., 2018. The role of prosody for the interpretation of rhetorical questions in German. In: Proceedings of the 9thInternational Conference on Speech Prosody 2018. Poznań, Poland, pp. 192–196.
Niebuhr, O., Winkler, J., 2017. The relative cueing power of f0 and duration in German prominence perception. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech). Stockholm, Sweden, pp. 611–615.
Nguyen, V.L., Shaker, M.H., Hüllermeier, E., 2022. How to measure uncertainty in uncertainty sampling for active learning. Mach. Learn 111 (1), 89–122.
Pardo, J.S., 2006. On phonetic convergence during conversational interaction. J. Acoust. Soc. Am. 119 (4), 2382. https://go.exlibris.link/PpJ6Wgs9.
Petrone, C., Niebuhr, O., 2014. On the intonation of german intonation questions: the role of the prenuclear region. Lang. Speech 57 (1), 108–146. https://doi.org/10.1177/0023830913495651.
Petrone, C., Truckenbrodt, H., Wellmann, C., Holzgrefe-Lang, J., Wartenburger, I., Höhle, B., 2017. Prosodic boundary cues in German: evidence from the production and perception of bracketed lists. J. Phon. 61, 71–92. https://doi.org/10.1016/j.wocn.2017.01.002.
Schertz, J., Clare, E.J, 2020. Phonetic cue weighting in perception and production. WIREs Cognit. Sci. 11 (2), e1521. https://doi.org/10.1002/wcs.1521.
Settles, B., 2009. Active Learning Literature Survey. Computer Sciences Technical Report at the University of Wisconsin-Madison.
Settles, B., 2012. Active Learning. Synthesis Lectures On Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers.
Sevastjanova, R., El-Assady, M., Hautli-Janisz, A., Kalouli, A.L., Kehlbeck, R., Deussen, O., Keim, D.A., Butt, M., 2018. Mixed-initiative active learning for generating linguistic insights in question classification. In: Proceedings of the 3rd Workshop on Data Systems for Interactive Analysis (DSIA) at IEEE VIS. Berlin.
Shen, D., Zhang, J., Su, J., Zhou, G., Tan, C.-L., 2004. Multi-Criteria-based Active Learning for Named Entity Recognition. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pages 589–596, Barcelona, Spain.
Shiamizadeh, Z., Caspers, J., Schiller, N.O., 2017. The role of F0 and duration in the identification of wh-*in-situ* questions in Persian. Speech Commun. 93, 11–19. https://doi.org/10.1016/j.specom.2017.07.005.
Stevens, K.N., Klatt, D.H., 1974. Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Am. 55, 653–659.
Thompson, C., Califf, M., Mooney, R., 1999. Active learning for natural language parsing and information extraction. In: Proceedings of the 16th International Conference on Machine Learning (ICML-99), pp. 406–414.
van Heuven, V.J., de Jonge, M., 2011. Spectral and temporal reduction as stress cues in Dutch. Phonetica 68 (3), 120–132. https://doi.org/10.1159/000329900.
Vendrig, D.H.V.J., Hartog, J., Leeuwen, D., Patras, I., Raaijmakers, S., Rest, J., Snoek, C., Worring, M., 2002. TREC feature extraction by active learning. In: Proceedings of the 11th Text Retrieval Conference (TREC).
Wagner, P., Cwiek, A., Samlowski, B., 2016. Beat it! Gesture-based prominence annotation as a window to individual prosody processing strategies. In: Proceedings of the Persentation at the 12th Conference on Phonetics and Phonology in German-speaking countries (PandP). München, Germany.
Wooldrigde, J.M., 2016. Introductory Econometrics: A Modern Approach. Cengage, Boston, MA.
Wu, Y., Kozintsev, I., Bouguet, J.Y., Dulong, C., 2006. Sampling strategies for active learning in personal photo retrieval. In: Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2006. Toronto, Ontario, Canada, pp. 529–532.
Zahner, K., Kutscheid, S., Braun, B., 2019. Alignment of f0 peak in different pitch accent types affects perception of metrical stress. J. Phon. 74, 75–95. https://doi.org/10.1016/j.wocn.2019.02.004.
Zhang, W., Liao, H., Zhao, N., 2008. Research on the FP growth algorithm about association rule mining. In: Proceedings of the International Seminar on Business and Information Management 2008, 1, pp. 15–318.
Zhang, X., 2012. A Comparison of Cue-Weighting in the Perception of Prosodic Phrase Boundaries in English and Chinese PhD Thesis. The University of Michigan.