Towards Relatable Explainable AI with the Perceptual Process

Wencan Zhang
National University of Singapore
Singapore
wencanz@u.nus.edu

Brian Y. Lim
National University of Singapore
Singapore
brianlim@comp.nus.edu.sg

ABSTRACT
Machine learning models need to provide contrastive explanations, since people often seek to understand why a puzzling prediction occurred instead of some expected outcome. Current contrastive explanations are rudimentary comparisons between examples or raw features, which remain difficult to interpret, since they lack semantic meaning. We argue that explanations must be more relatable to other concepts, hypotheticals, and associations. Inspired by the perceptual process from cognitive psychology, we propose the XAI Perceptual Processing Framework and RexNet model for relatable explainable AI with Contrastive Saliency, Counterfactual Synthetic, and Contrastive Cues explanations. We investigated the application of vocal emotion recognition, and implemented a modular multi-task deep neural network to predict and explain emotions from speech. From think-aloud and controlled studies, we found that counterfactual explanations were useful and further enhanced with semantic cues, but not saliency explanations. This work provides insights into providing and evaluating relatable contrastive explainable AI for perception applications.

CCS CONCEPTS
• Human-centered computing → Empirical studies in HCI. • Computing methodologies → Artificial intelligence.

KEYWORDS
Explainable AI, contrastive explanations, audio, vocal emotion

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ACM Reference format:
Wencan Zhang and Brian Y. Lim. 2022. Towards Relatable Explainable AI with the Perceptual Process. In CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA. ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3491102.3501826

1 INTRODUCTION
With the increasing availability of data, deep learning-based artificial intelligence (AI) has achieved strong capabilities in computer vision [50], natural language processing [45], and speech processing [5]. However, their complexity limits their use in real-world applications due to the difficulty to understand them [31]. To address this, much research has been conducted on explainable AI (XAI) to develop new XAI algorithms and techniques [3, 7, 30, 36], understand user needs [22, 58, 59, 72, 104] and evaluate their helpfulness [2, 11, 12, 18, 21, 64, 83, 105].

Despite the myriad XAI techniques, many of them remain difficult to understand, due to the lack of human-centered design that do not satisfy human needs [23, 72, 104]. Miller identified contrastive reasoning as a particular reason that people ask for explanations [72] — "one does not explain events per se, but that one explains why the puzzling event occurred in the target cases but not in some counterfactual contrast case." [35]. We further argue that explanations lack relatability towards concepts that people are familiar with, and therefore they seem too low-level technical and not semantically meaningful. Existing contrastive explanation techniques [20, 54, 73, 102] remain unrelatable, hence limiting their interpretability. In this work, we extend the framing of relatable explanations beyond contrastive explanations to include saliency, counterfactuals, and cues. Explanations should be relatable towards concepts via contrastive explanations, towards exemplars by providing counterfactual examples, and towards associated auxiliary concepts such as sensory and semantic cues.

We have identified audio prediction as a problem space in dire need of relatable explanations. Much research on XAI techniques focuses on structured data with semantically meaningful features, unstructured data such as text with semantically meaningful words or sentences, and images with visualizations that are visually intuitive. Explaining audio visually is problematic, since sound is not visual and people understand them through relating to concepts or other audio samples [79]. Current explanation techniques for audio typically present saliency maps on audiograms or spectrograms. Spectrograms are too technical for lay users or even non-engineering domain experts. Saliency maps are too simple and merely point to regions without explaining their relevance. Example-based explanations extract or produce examples for users to compare, but this still requires people to speculate why some examples are similar or different. Hence, explaining audio predictions requires relating the prediction to other concepts, counterfactual examples, and associated cues. We study the use case of vocal emotion recognition to propose relatable explanations. With applications in smart speakers for the home [70], digital assistants for mental health monitoring [9, 106], and affective computing [81], there is a growing need for these AI models to be relatably explainable.

Furthermore, not only should explanations be semantically meaningful, but the way the explanations are generated or the way the AI "thinks" should be human-like to earn people’s trust [93]. We draw inspiration from theories of human cognition to understand why and how people relate concepts, information, and data. Specifically, we frame relatable explanations with the Perceptual Process [13], where people select, organize, and interpret information to make a decision. Corresponding to these stages, we propose the XAI Perceptual Processing Framework with modular explanations for
Contrastive Saliency, Cues, and Counterfactual Synthetics with Contrastive Cues, respectively. This was implemented as RexNet (Relatable Explanation Network), a deep learning model with modules for each explanation type. We evaluated the explanations with a modeling study, a qualitative think-aloud study and a quantitative controlled study to investigate their usage and impact on decision performance and trust perceptions. We found that RexNet improved prediction performance and explanation faithfulness; participants appreciated the diversity of explanations; and participants benefited from Counterfactual and Cues explanations, but not for Saliency explanations. In summary, we address the challenge that explanations need to be relatable, and studied this for an audio prediction task (vocal emotion recognition). Our contributions are:

(1) RexNet Perceptual Processing Framework for relatable explanations inspired from theories in human cognition.
(2) RexNet model with multiple relatable explanation (Contrastive Saliency, Counterfactual Synthetic, Contrastive Cues).
(3) First to provide relatable explanations for audio predictions.
(4) Evaluation of usage and usefulness of relatable explanations.

2 RELATED WORK
We introduce various explainable AI techniques, argue how they lack human-centeredness, and describe the background on speech emotion recognition and highlight their lack of explainability.

2.1 Explainable AI techniques
Much research has been done to develop explainable AI (XAI) for improving model transparency and trustworthiness. An intuitive approach is to point out which features are most important. Attribution explanations do this by identifying importance using gradients [99], ablation [92], activations [97], or decompositions [8, 75, 90]. In computer vision, attributions take the form of saliency maps (e.g., [97]). Explaining by referring to key examples is another popular approach. This includes simply providing arbitrary samples of specific classes, cluster prototypes or criticisms [46], or influential training set instances [49]. However, users typically have expectations and goals when asking for explanations.

Users ask for contrastive explanations when expected outcomes do not happen. A simple answer would find the attribution differences between the actual (fact) and expected (foil) outcomes [84] However, this is naive because users are truly asking for what differences in feature values, not attributions, would lead to the alternative outcome. That is a counterfactual explanation. Furthermore, to anticipate a future outcome or prevent an undesirable one, users could ask for counterfactual explanations. Indeed, contrastive explanations are often conflated with counterfactual explanations in the research literature. Such explanations suggest the minimum changes in the current case to achieve the desired outcome [103]. Trained decision structures, such as local foil trees [102], Bayesian rule lists [55], or structural causal models [73] can also serve as counterfactual explanations. Though typically explained in terms of feature values [20, 54, 103] or anchor rules [91], techniques have been developed to synthesize counterfactuals of unstructured data (e.g., images [29] and text [33]). In this work, we employ the synthesis approach to generate counterfactuals of audio data.

There are many explanation types and Lim and Dey have framed them in an intelligibility taxonomy as Why (Attribution), Why Not (Contrastive), and How To (Counterfactual) [59]. Many of these XAI techniques have been independently developed or tested, so their usage is disparate. In this work, we unify them in a common framework and integrate them in a single machine learning model.

2.2 Human-Centered Explainable AI
Abdul et al. [1] found a large gap between XAI algorithms and human-centered research. To close this gap, HCI researchers have been active in evaluating the various benefits of XAI or lack thereof, including understanding and trust [64], uncertainty [62, 105, 110], cognitive load [2], types of examples [11], etc. Studies have sought to determine the “best” explanation type [64, 100], but others have revealed the benefit of reasoning with multiple explanations [6, 61, 63]. Hence, we propose a unified framework to provide multiple relatable explanations together. We determined our human-centered explanation requirements by studying literature on human cognition, which is epistemologically similar to works grounded in philosophy and psychology [72, 104], and unlike empirical approaches to elicit user requirements [22, 58, 59]. Furthermore, current works focus on explaining higher-level reasoning tasks, but not perception tasks that are commonplace. This has implications on the depth of explanations to provide, which we investigate in this work.

2.3 Speech Emotion Recognition
Deep learning approaches proliferate research on automatic speech emotion recognition (SER). Leveraging the intrinsic time-series structure of speech data, recurrent neural network (RNN) models with attention mechanism have been developed to capture transient acoustic features to understand contextual information [74]. Employing popular techniques from the computer vision domain, audio data can be treated as 1D arrays or converted to a spectrogram as a 2D image. Convolutional neural networks (CNNs) can then extract spatial features from these audiograms or spectrograms [37]. Current approaches improve performance by combining CNN and RNN [101, 114], or modeling with multiple modalities [111]. Our RexNet model starts with a base CNN model to leverage many more XAI techniques available to CNNs than RNNs. Since our approach is modular, it can be generalized to state-of-the-art SER models.

2.4 Model Explanations of Audio Predictions
Due to the availability of image data and intuitiveness of vision, much XAI research has focused on image prediction tasks; in contrast, few techniques have been developed for audio prediction tasks. Many techniques exploit CNN explanations by generating a saliency map on the audio spectrogram [4, 52]. Other explanations focus on model debugging by visualizing neuron activations [51], or as feature visualizing [56] (like [78] for image kernels). We also leverage saliency maps as one explanation, due to its intuitive pointing, but augment it with relatable explanations. Other than explaining the model behavior post-hoc, another approach is to make the model more interpretable and trustworthy by constraining the trained model with domain knowledge, such as with voice-specific parametric convolutional filters [67, 88]. Our approach with modular explanations of specific types follows a similar objective.
Stimuli
"dog face" and "cat face" by "irfan al haq", "Dog" by Maxim Kulikov, "cat mouth" by needumee from the Noun Project.

We discuss how the framework supports relatable explanations, and with the smallest difference. For our application in vocal emotion whether each element is closer to the cat or dog version (Fig. 1 uses wards higher-level concepts. In our example, the face cues are used (to synthesize counterfactuals, compare cues, classify concepts). For visual clarity, we present the use case for visually recog-

The perceptual process defines three stages for how humans per-
ception and cognition to define our explainable AI techniques. We discuss how the framework supports relatable explanations, and apply it to vocal emotion recognition. Next, we describe background theories from cognitive psychology and vocal emotion prosody, and define requirements for relatable explanations.

3 INTUITION AND BACKGROUND
To improve trust, models should provide explanations that are rel-
able and human-like. Thus, we propose to use theories of human perception and cognition to define our explainable AI techniques. We discuss how the framework supports relatable explanations, and apply it to vocal emotion recognition. Next, we describe background theories from cognitive psychology and vocal emotion prosody, and define requirements for relatable explanations.

3.1 Perceptual Processing
The perceptual process defines three stages for how humans per-
ceive and understand stimuli: selection, organization, and interpre-
tation [13]. Fig. 1 illustrates these stages for the case of visually perceiving a cat and relates them to our technical approach. When sensory stimuli (e.g., light rays or audio vibrations) reach the senses, 1) our brain first selects only a subset of the information to focus attention. This is equivalent to highlighting salient regions in an image. 2) The next stage organizes the salient regions into meaningful cues. For a face, these would include recognizing the ears, eyes, and nose. 3) Finally, the brain interprets these lower-level cues towards higher-level concepts. In our example, the face cues are used to recognize the animal by: a) recalling from long-term memory the concepts of cat and dog, and their respective cues, b) compare whether each element is closer to the cat or dog version (Fig. 1 uses a slider paradigm for illustration), and c) categorize the concept with the smallest difference. For our application in vocal emotion recognition, the perceptual processing framework aligns with Kotz et al’s model for processing emotion prosody [79, 95] that describes stages for "extracting sensory/acoustic features, detecting meaningful relations, conceptual processing of the acoustic patterns in relation to emotion-related knowledge held in long-term memory."

In particular, people categorize concepts by mentally recalling examples and comparing their similarities [27]. These examples may be prototypes or exemplars. With Prototype Theory, people summarize and recall average examples, but these may be quite different from the observed case being compared. With Exemplar Theory, people memorize and recall specific examples, but this does not scale with inexperienced cases. Instead, people can imagine new cases that they have never experienced [10]. Moreover, rather than tacitly comparing some ill-defined difference between the examples, people make comparisons by judging similarities or differences along dimensions (cues) [77]. Categorization can then be done systematically with proposition rules or intuitively [40], with either sometimes being more effective [69].

We apply this framework and propose a unified technical ap-
proach with contrastive explanation types to align with each stage of perceptual processing: 1) highlight saliency, 2) recognize cues, 3a) synthesize counterfactual, 3b) compare cues, and 3c) classify concept. We further present cue differences as rules and leverage an embedding for emotions to represent intuition (described later).

3.2 Desiderata for Relatable Explanations
Informed by the Perceptual Process, Prototype and Exemplar Theo-
ries, we identified requirements that AI explanations of the predic-
tion of an instance should be made more relatable towards:

- **Concepts** by relating the predicted concept to other concepts. Contrastive explanations [59, 72] are thus a key foundation for broader relatable explanations.
- **Exemplars** by comparing the factual (actual) instance with counterfactual instances of the other concepts. Concepts are abstract, so providing concrete examples can help people to fixate on details and cues for comparison. Counterfactual explanations [20, 72, 103] are a first step in identifying marginally different instances with different prediction outcomes, but do not further relate to why the instances are different.
- **Cues** by relating how auxiliary concepts or associated cues are different between the factual and counterfactual instances. For perception tasks, this involves highlighting saliency in sensory cues (stimuli; e.g., eyes of a face). For cognition tasks, this involves articulating differences in semantic cues (e.g., interpreted speech rate from phonemes). Attribute value explanations are
We trained a vocal emotion classifier on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset [66] with 7356 audio clips of 24 voice actors (50% female) reading fixed sentences with 8 emotions (neutral, calm, happy, fearful, surprised, sad, disgust, angry). Each audio clip was 2.5–3.5 seconds long, and we padded or cropped them to a fixed 3.0s. We parsed each audio file to a time-series array of 48k readings (i.e., 16 kHz sampling rate), and preprocessed it to obtain a mel-frequency spectrogram with 128 frequency bins, 0.04s window size, and 0.01s overlap. Treating the spectrogram as a 2D image, we can train a convolutional neural network (CNN) [34]. Specifically, we trained a CNN with 3 convolutional blocks, and 2 fully connected layers. We used cross-entropy loss for multi-class classification. In sum, the base CNN model \( M_0 \) takes audio input \( x \) to predict an emotion \( y_0 \) (lower left in Fig. 2).

### 4 TECHNICAL APPROACH

We propose an interpretable deep neural network to predict vocal emotions and provide relatable explanations. We first describe the base prediction model, then specific explanation modules.

#### 4.1 Base Prediction Model

We trained a vocal emotion classifier on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset [66] with 7356 audio clips of 24 voice actors (50% female) reading fixed sentences with 8 emotions (neutral, calm, happy, fearful, surprised, sad, disgust, angry). Each audio clip was 2.5–3.5 seconds long, and

---

**Table 1: Vocal cues for emotion recognition.**

| Vocal Cue [39]       | Simple Name  | Description                                                                 |
|----------------------|--------------|-----------------------------------------------------------------------------|
| High-Frequency Energy (HF 500) | Shrillness | Proportion of high-frequency energy (cut-off 500 Hz) in the acoustic spectrum, i.e., how much of the speech is high-pitch. |
| Voice Intensity    | Loudness    | Mean of sound amplitude, i.e., how loud the person is speaking.             |
| Mean \( F_0 \)     | Average Pitch | Mean of fundamental frequency, i.e., pitch.                                |
| SD \( F_0 \)       | Pitch Range  | Standard deviation of fundamental frequency, i.e., pitch variation.         |
| Speech Rate        | Speaking Rate | How quickly the person is speaking (words/second), i.e., \( 1/t_{total} \). |
| Pause Proportion   | Proportion of Pauses | Proportion of pauses in the speech, i.e., \( t_{pauses}/t_{total} \). |

---

We extended the base model with multiple modules to provide three relatable contrastive explanations (Fig. 2). The whole architecture can be understood in terms of a chain of dependencies. We describe this in reverse starting with the goal. Ultimately, we want to explain the prediction with descriptive contrastive cues. This requires a counterfactual 'foil' to compare the target ‘fact’ with, therefore, we need to obtain an example for comparison. When making a comparison, not all stimuli are relevant for interpretation; hence, we need to select salient segments. For example, noticing a flower in a photo of a cat is irrelevant to identifying whether an animal is a dog or cat. In summary, our approach has steps:

1. Highlight salient segments
   - i. Predict emotion concept as initial estimation
   - ii. Keep embedding vector of estimation for final classification
   - iii. Explain contrastive saliency using discounted Grad-CAM
2. Describe segments
   - i. Infer associated cues
3a. Generate counterfactual exemplar for each contrast concept
   - i. Generate counterfactual synthetic using StarGAN-VC [41]
3b. Compare cue differences between target case and each exemplar
   - i. Calculate cue differences weighted by saliency
   - ii. Classify cue difference relations with cue differences and embedding for target and contrast concepts.
3c. Classify concept fully
   - i. Predict concept using inputs: cue differences of all counterfactuals + embedding (initial estimation)
   - ii. Explain final concept with attributes for cue differences using Layer-wise Relevance Propagation (LRP) [8]

We next describe each module for specific contrastive explanations.

#### 4.2 RexNet: Relatable Explanation Network

We introduce RexNet — Relatable Explanation Network\(^1\) — to provide relatable explanations for contrastive explainable AI. People recognize vocal emotions based on various vocal stimulus types and prosodic attributes [53], such as verbal [39] and non-verbal expressions [94] (e.g., laughs, sobs, screams), and lexical [71] information. In this work, we focus on vocal cues (prosody) identified by Juslin et al. (e.g., see Table 1). These cues are about how words are spoken, rather than the words themselves (lexical information). We leverage people’s ability to index vocal emotion categories by the pattern of cues [39] to identify cue differences between different emotions, which we present in our model explanation. Although people may be able to perceive various vocal cues, they may be unable to relate to them conceptually (e.g., ‘formant frequency’ is technically complex), therefore, we limit cues to familiar everyday concepts. In our user study, we further verified their understandability in a screening test. For our prediction application, the concept to predict is emotion, cues are vocal cues for emotion prosody, cue differences support dimensional comparisons, and saliency is in terms of phonemes or pauses between them.

\(^1\)The ‘s’ can also stand for a cross for Why Not to indicate contrastive explanations.
Figure 2: Modular architecture of RexNet with relatable explanations for the prediction of emotion $y$ from input voice $x$. Each module is numbered to match the sequence of the perceptual process (Fig. 1). Black arrows indicate feedforward activations. Red arrows indicate backpropagation during training. The base CNN model $M$ is denoted as a trapezium block to represent its function as an encoder. The StarGAN generator $G_{GAN}$, represented as an encoder-decoder, takes input $x$ and output $\hat{x}^y$ with the same shape. $M_c$ is a heuristic model, and $M_r$ and $M_y$ are sub-models with only fully-connected layers. Although we trained the model on 2D spectrograms, for illustrative simplicity, the audio data is represented in its 1D audio waveform.

from faces may always highlight the eyes regardless of emotion. To address the issue of saliency lacking semantic meaningfulness, we introduce associative cues, which we describe later. Here, we address the need for more specific saliency with a discounted saliency map to produce contrastive saliency. This retains some importance of globally important pixels, unlike current methods that simply subtract a saliency map of one class from that of another class [84]. Dhurandhar et al. [20] identified pertinent positives and negatives for more precise contrastive explanations by perturbing features, but our approach calculates based on feature activations.

We define two forms of contrastive saliency: pairwise and total. Pairwise contrastive saliency highlights pixels that are important for predicting target class $y$ but discounts pixels that are also important for foil class $\gamma$. We implemented the saliency map with Grad-CAM [97], and define the class activation map for class $y$ as $s^y$. The pairwise contrastive saliency between classes $y$ and $\gamma$ is thus:

$$s^{y\gamma} = \lambda^{y\gamma} \odot s^y$$  \hspace{1cm} (1)$$

where $\lambda^{y\gamma} = (1 - s^\gamma)$ indicates the discount factors for all pixels due to their attributes to class $\gamma$. 1 is a matrix of all ones, and $\odot$ is the Hadamard operator for pixel-wise multiplication. To identify important pixels for class $y$ but not any other class, we define total contrastive saliency as:

$$s^y = \lambda^y \odot s^y$$  \hspace{1cm} (2)$$

where $\lambda^y = \sum_{\gamma \in C \setminus \{y\}} (1 - s^\gamma) / |C - 1|$ indicates the discount factors across all alternative classes, and $C$ is the number of classes.

In RexNet, the saliency explanation is calculated from the initial emotion classifier $M_0$ predicting an initial emotion concept $\hat{y}_0$. We present contrastive saliency for audio using a 1D saliency bar aligned to words in the speech (see Fig. 5), which aggregates saliency in the spectrogram across frequencies per time bin. This is more accessible for lay people to understand since it avoids using technical spectrograms or audiograms (audio waveforms).

4.2.2 Contrastive Synthetic. Due to the open-ended variability in unstructured data, contrastive samples drawn from a training set are likely to be quite different from the target instance. Contrastive samples will have extraneous differences that may be distracting to interpret and less meaningful for comparison. Instead, contrastive synthetics are generated to be similar to the target instance, except sufficient differences to achieve the contrastive outcome. Fig. 3 illustrates the benefit of using contrastive synthetics for comparison. When deciding whether a target item is more similar to a first or second reference, one would measure the target’s distance to each reference. Contrastive synthesis produces comparison references that are closer to the target item being classified, because it minimizes the differences between the target item and reference example. These contrastive synthetics will be closer to other model samples that the model knows (prior instances in the training set), model prototypes (centroids or medoids of class clusters), or human mental exemplars (from the user’s memory), since the model may not have a similar example or the human may never have seen or heard a very similar case to the target item. This amplifies the ratio between the reference distances larger, and makes the difference more perceptible. Formally, the ratio of differences for contrastive synthetics are larger than for other examples (prototypes, or samples of prior items), i.e., $|\log(\delta_1/\delta_2)| > |\log(d_1/d_2)|$. Therefore, contrastive synthetics help make comparison between references more easy.

We aim to create a contrastual that is similar to the target instance $x$ which is classified as class $y$, but with sufficient differences to be classified as another class $\gamma$. Current counterfactual methods focus on structured (tabular) data by minimizing changes to the target instance [76, 103], or identifying anchor rules [91], but this is not possible for unstructured data (e.g., images, sounds). Instead, inspired by data synthesis with Generative Adversarial Networks (GANs) [44, 86] and style transfer [16, 117], we propose...
As a generative adversarial model, StarGAN trains three models—a generator $G$, discriminator $D$, and domain classifier $M$. $G$ inputs the target instance $x$ that is of class $y$ and the objective class $\hat{y}$ to generate a similar instance $x'$. The training objectives are to make $x' \approx x$ and $M(x') \approx \hat{y}$. Next, $x'$ and $y$ are input into $G$ to get $\hat{x}'$ as output. $G$ is trained to minimize the cycle consistency reconstruction loss between $x$ and $\hat{x}'$, which also improves $x'$. $\hat{x}'$ is also input into the $M$ to output class $\hat{y}$, which is trained to minimize the loss between $\hat{y}$ and $\hat{y}$. Finally, $D$ is trained to ensure that the generated instances are more realistic $\hat{d}$. Together, this semi-supervised method trains $G$ to generate style-transferred instances.

### 4.2.3 Contrastive Cues

The final contrastive explanation involves first inferring cues from the target and counterfactual instances and comparing them. We define the individual cue as absolute cues ($c^y$ and $\hat{c}^\hat{y}$), and the difference as contrastive cues $\delta_{c}^y$. We report on 6 vocal cues identified by Juslin and Laukka [39] for vocal emotions (Table 1). Absolute cues can be inferred with machine learning predictions or heuristically. For vocal emotions, since cues can be deterministically measured from the input data, we use heuristic methods $M_c$ to infer the cues $c$. For example, pitch range is calculated as follows: a) calculate fundamental frequency (modal frequency bin) for each CAM-salient time window in the spectrogram, b) calculate their standard deviation. For semantically abstract cues, such as sounding “melodic”, “questioning”, or “nasally”, they should be annotated by humans and inferred using supervised learning.

We calculated contrastive cues as ordinal cue difference relations $\delta_{c}^y$ from numeric cue differences $\delta_{c}^\hat{y}$ based on the instances in the RAVDESS dataset [66]. To determine differences between emotions for each cue, we fit the data to a linear mixed effects model with emotion as the main fixed effect and voice actors as random effect (see Supplementary Fig. 1), and performed a Tukey HSD test with significance level $\alpha = .005$ to account for the multiple comparison effect. For each cue, if an emotion is not significantly higher than the other, then we label the cue difference as “similar”; otherwise, we label it as “higher” or “lower” depending on the direction. Table 2 describes the vocal cue patterns of each emotion compared to average levels, which is in close agreement with [39] except for the fearful emotion. Table 3 describes the pairwise cue difference relations between each emotion and an emotion (happy).

Predicting the cue difference relations $\delta_{c}^y$ requires deciding the decision threshold at which to split the cue difference $\delta_{c}^y$ to categorize the relation, and this can contextually depend on initially estimating which emotion concepts $\hat{y}_0$ and $y_0$ to compare, and which cues are more relevant. We define this as a multi-task model with two sub-models with fully connected neural network layers $M_c$ and $M_y$. $M_c$ takes in the numeric cue differences $\delta_{c}^\hat{y}$ and embedding representations (from the penultimate fully connected layer) of the emotion concepts $\hat{z}_0$ and $\hat{z}_0$ to predict the emotion $\hat{y}$ heard in $x$. We determine which cues were more important by calculating an attribution explanation $\hat{w}_c^y$ with layer-wise relevance propagation (LRP) [8]. These attributions are then concatenated on the $\delta_{c}^y$ to determine the weighted cue differences $w_{c}^y$. $M_y$ takes in $w_{c}^y$, $\hat{z}_0$, and $\hat{z}_0$ to predict the cue difference relations $\delta_{c}^y$. With the ground truth references, cue difference relations prediction can be trained using supervised learning. Since the cue difference relations (lower, similar, higher) are ordinal, we employed the NNRank ordinal encoding [15] with 2 classes, such that lower $= (0,0)^T$, similar $= (1,0)^T$, higher $= (1,1)^T$, sigmoid activation, and binary cross-entropy loss for multi-label classification.

### Table 2: Vocal cues for emotions relative to average levels.

| Target Emotion | Vocal Cue | Shrillness | Loudness | Average | Pitch | Range | Speaking Rate | Proportion of Pauses |
|----------------|----------|------------|----------|---------|-------|-------|---------------|---------------------|
| Neutral        |          | low        | low      | low     | high  | low   | low           | low                 |
| Calm           |          | low        | low      | low     | average | low | low           | average             |
| Happy          | high     | high       | high     | high    | high  | high  | high          | low                 |
| Fearful        |          | average    | average  | high    | high  | low   | high          | high                |
| Surprised      |          | low        | low      | low     | low   | low   | low           | average             |
| Sad            | low      | low        | low      | low     | low   | low   | low           | average             |
| Disgust        |          | average    | low      | high    | low   | low   | low           | high                |
| Angry          |          | high       | high     | high    | high  | low   | low           | average             |

### Table 3: Contrastive vocal cues for target emotions compared to another emotion (Happy).

| Target Emotion | Vocal Cue | Shrillness | Loudness | Average | Pitch | Range | Speaking Rate | Proportion of Pauses | Contrast Emotion |
|----------------|----------|------------|----------|---------|-------|-------|---------------|---------------------|------------------|
| Neutral        |          | low        | low      | low     | high  | low   | low           | low                 | Happy            |
| Calm           |          | low        | low      | low     | similar | low | low           | low                 | Happy            |
| Happy          | similar  | similar    | similar  | similar | similar | similar | similar       | similar             | Happy            |
| Fearful        |          | similar    | similar  | similar | similar | similar | similar       | similar             | Happy            |
| Surprised      |          | low        | low      | low     | low   | low   | low           | low                 | Happy            |
| Sad            | low      | low        | low      | low     | low   | high  | low           | high                | Happy            |
| Disgust        |          | average    | low      | high    | low   | low   | low           | high                | Happy            |
| Angry          |          | higher     | similar  | similar | similar | low   | similar       | similar             | Happy            |
We evaluated the faithfulness of our interpretable model, then the input metrics: 1) reconstruction similarity \( \text{MSE} \), calculated with mean square error

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2
\]

that identifies more important features as those that cause larger decreases in model performance when that feature is ablated. 

Towards Relatable Explainable AI with the Perceptual Process CHI ’22, April 29-May 5, 2022, New Orleans, LA, USA

Table 4: Evaluation results of model prediction performance and explanation correctness for RexNet with StarGAN or Counterfactual Samples and baseline models. Grey numbers calculated from definition. * same as Base CNN.

| Relation | Cue accuracy | Model |
|----------|--------------|-------|
| Cue Difference | |        |
| \( \hat{r}_{gf} \) | 71.9% | RexNet w/ C.Samples |
| \( \hat{r}_{gf} \) | 71.6% | RexNet w/ C.Samples |
| \( \hat{r}_{gf} \) | 71.3% | Base CNN |
| \( \hat{r}_{gf} \) | 71.2% | Random |

5.1 Modeling Study

Method. We evaluated the model prediction performance and explanation correctness with several metrics (Table 4). We measured the accuracies of the initial and final predictions of emotion, and compared them against that of the baseline CNN model. Each explanation type was evaluated with different metrics due to their different forms. We evaluated saliency maps by the relevance of important features to the model prediction, and compared absolute and contrastive saliency. We employed the ablation approach of [57] that identifies more important features as those that cause larger decreases in model performance when that feature is ablated. We evaluated the faithfulness of counterfactual synthetics with these metrics: 1) reconstruction similarity \( \exp(-\text{MSE}(x, \hat{x}_p)) \) between the input \( x \) and synthesized \( \hat{x}_p \), calculated with mean square error

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2
\]
rather than investigating whether each explanation as implemented is good enough. Therefore, we select instances with correct predictions and coherent explanations for the user studies. Since the Counterfactual Synthesis performance is limited, we use Counterfactual Samples to represent counterfactual examples instead.

5.2 Think-Aloud User Study
We conducted a formative study with the think-aloud protocol to understand how people 1) naturally infer emotions without AI assistance, and 2) use or misunderstand various explanations.

5.2.1 Experiment Method and Procedure. We recruited 14 participants from a university mailing list. They were 3 males, 11 females, with ages between 21-40 years old. We conducted the study via an online Zoom audio call. The experiment took 40-50 minutes and each participant was compensated with a $7.43 USD coffee gift card. The user task is a human-AI collaborative task for vocal emotion recognition. Given a voice clip, the participant infers the portrayed emotion with or without AI prediction and explanation. We provided 16 voice clips of 2 neutral sentences\(^2\) intoned to portray 8 emotions. We selected only correct system predictions and explanations, since we were not investigating the impact of erroneous predictions or misleading explanations. The study contains 4 explanation interface conditions: Contrastive Salience only, Counterfactual Sample voice examples only, Counterfactual Sample and Contrastive Cues, and all 3 explanations together (Fig. 5).

The procedure is: read an introduction, consent to the study, complete a guided tutorial of all explanations (regardless of condition), and start the main study with multiple trials of a vocal emotion recognition task. To limit the participation duration, each participant completes three trials, each trial randomly assigned to an explanation interface condition. For each trial, the participant listened to a voice clip, and gave an initial label of the emotion. On the next page, the participant was shown the system’s prediction with (or without) explanation based on the assigned condition. She could then revise her emotion label if she changed her mind. We used the think-aloud protocol to ask participants to articulate their thoughts as they examined the audio clip, prediction and explanations. We also asked them about their perceptions using the interface, and any suggestions for improvement. We describe our findings next.

5.2.2 Findings. We performed thematic analysis on the recorded audio to determine key themes (bolded). We describe our findings in terms of our research questions of how users innately infer vocal emotions, and how they used each explanation type. When inferring on their own (without XAI), participants would focus on specific cues to “check the intonations [pitch variation] for decision” [Participant P12], infer a Sad emotion based on the “flatness of the voice” [P04], or “use shrillness to distinguish between fearful and surprise” [P01]. Participants also relied on changes in tone, which we had not modeled. For example, a rising tone “sounds like the man is asking a question” [P02], “the last word has a questioning tone” [P03] helped participants to infer Surprise. The latter case identified the most relevant segment. In contrast, a “tone going down at the end of sentence” helped P01 infer Sad. Some participants mentally generated their own examples to “imagine what neutral sound like and compare against it” [P05]. These unprompted behaviors suggest the relevance of saliency, counterfactual, and cue explanations.

The usage of explanations was mixed with some benefits and some issues. In general, participants could understand the Saliency maps. P09 saw that “the highlight parts are consistent with my judgment for important words”, referring to ‘talking’ being highlighted. However, several participants had issues with saliency maps. There were some cases with highlights that spanned across multiple words and included highlighting spaces. P08 felt that saliency “should highlight all words”, and P14 “would prefer the color highlighted on text”. This lack of focus made P13 feel that “the color bar is not necessary”. Regularizing the explanation to prioritize highlighting words and penalize highlighting spaces can help align the explanations with user expectations and improve trust [93]. Next, P11 thought that “the color bar reflects the fluctuation of tone”. While plausible, this indicates the risk of misinterpreting technical visualizations for explanations. Finally, P12 “used the saliency bar by listening to the highlighted part of the words, and try to infer based on intonation. But I think the highlighting in this example is not accurate”. This demonstrates causal oversimplification by reasoning with one factor rather than multiple factors [19, 61].

Many participants found Counterfactual samples “intuitive”. P11 could “check whether it’s consistent with my intuition” by mentally comparing the similarity of the target audio clip (sad) with clips for other suspected emotions (neutral, sad, happy). Unfortunately, her intuition was somewhat flawed, since she inferred Neutral which was wrong. P12 found counterfactuals “helpful to have a reference state, then I will also check the intonations for my decision.” Conversely, some participants felt counterfactual samples were not helpful. P06 felt that the “clips [neutral and calm] are too similar”. Had she received deeper explanations with saliency map or cue differences, she would have had more information about where and what the differences were, respectively.

Cues were used to check semantic consistency. P04 used cues to “confirm my judgment” and found that the “low shrillness [of Sad] is consistent with my understanding.” However, some participants perceived inconsistencies. P13 thought that “some cue descriptions were not consistent with my perception”, and disagreed with the system that Speaking Rate was similar for the Happy and Surprised audio clips. Along with the earlier case of P06, this suggests differences in perceptual acuity of cues between the user and system.

Finally, some participants felt that Counterfactual samples were more useful than Contrastive Cues. P11 found that “the comparison voice part is more helpful than the text part, though the text part is also helpful to reinforce my decision.” This could be due to cognitive load and differences between mental dual processing [40]. Many participants considered the audio samples “quite intuitive” [P04]. They used System 1 thinking which is fast, though they did not articulate why this was simple. In contrast, they found that “it’s hard to describe or understand the voice cue patterns” [P04]. P10 felt that “compared with [audio] clips, cue pattern is too abstract to use for comparison.” This requires slower System 2 thinking. Another possible reason is that the audio clip has higher information bandwidth than the 6 verbally presented semantic cues. Participants can perceive the gestalt [48] of the audio to make their inferences.
5.3 Controlled User Study

Having identified various benefits and usages of contrastive explanation, we next conducted a summative controlled study to understand: 1) how well participants could infer vocal emotions on their own, and with model predictions and explanations, and 2) how various explanations affect perceived helpfulness.

5.3.1 Experiment Design and Apparatus. We conducted a between-subjects experiment with XAI Type as the independent variable with 5 levels of explanations (None, Contrastive Saliency, Counterfactual Sample, Counterfactual + Contrastive Cues, and Saliency + Counterfactual + Cues). The user task is to label the portrayed emotion in a voice clip with feedback from the AI in one of the XAI Types. We included emotion as a random variable with 8 levels. Having many emotions helps to make the task more challenging to test. Fig. 5 shows the UI with all explanations together, and others are shown in Supplementary Figs. 7-11. For dependent variables, we measured decision quality (emotion label correctness and confidence), understanding of cue differences, task times, decision confidence, and perceived system helpfulness. Labeling correctness was measured with a “balls and bins” question [26] that elicits the probability of multiple labels. Cue difference understanding was measured per cue with a multiple choice question for the cue difference relation between a randomly selected contrast emotion label and the target voice clip. Task times were logged for different pages. Perceptions were measured as ratings on a 7-point Likert scale (−3 = Strongly Disagree, 4+ = Strongly Agree). We asked two text questions about the rationale for perceived helpfulness and how the explanation was used. This was posed only twice to limit fatigue. See Supplementary Figs. 7-12 for the survey.

5.3.2 Experiment Procedure. The participant reads an introduction, consents to the study, reads a short tutorial about the explanation interfaces, and completes a screening test of audio equipment, auditory acuity, and UI understanding (Supplementary Figs. 2-4), where she: a) listens to a voice clip and chooses the correct spoken words, b) reads a saliency map and identifies important words, and c) identifies easy cue differences between two voice clips.

After passing screening (with all questions correct), the participant is randomly assigned to an XAI Type and commences a practice session. Similar to [64], we conducted the practice session to enable the participant to learn from any model explanations how the system predicts the emotion. She is encouraged to study these cases carefully, since she will not see the correct predictions later in the main study. The practice session comprises 8 trials, where each trial has three pages: i) Pre-AI to listen to a voice clip, label the emotion without AI assistance. We assess the labeling correctness here to estimate the Participant Unaided Skill, i.e., whether the participant has above- or below-average skill in recognizing vocal emotions. ii) Post-XAI to read any explanation feedback (without seeing the system prediction), label the emotion (again), and answer questions about cue difference understanding, and perceived ratings. iii) Review to examine the correct emotion label (same as the system prediction) with any AI explanations, and the participant’s previous answer, and write any free-form notes (open text).

After the practice session, the participant engages in the main study with the same XAI Type in two sessions with 8 trials each and a break in-between. Each trial is presented on one page where the participant: i) listens to the voice clip, ii) views any explanation feedback, iii) labels the emotion, and iv) rates perceptions. This evaluates human-simulatability [64, 65] by deeply testing the participant’s understanding to apply explanations to new instances. To control any fatigue effects due to the moderate number of trials, we randomized the order of instances. We asked the rationale questions randomly in one trial per main session. The participant is incentivized to be fast and correct with a maximum $0.50 USD bonus for completing all trials within 8 minutes. The bonus is prorated by the number of correct emotion labels. Maximum bonus is $1.00 for two sessions over a base compensation of $3.00 USD. The participant ends with answering demographic questions.

5.3.3 Statistical Analysis and Quantitative Results. We recruited 175 participants from Amazon Mechanical Turk with high qualifications (≥ 5000 completed HITs with >97% approval rate). They were 52.0% male, with ages 21-70 (Median = 36). Participants took 27.4 minutes (median) to complete the survey. We excluded 14 participants who completed the survey without playing any voice clips. For each dependent variable, we fit a linear mixed-effects model with XAI Type, Emotion, Voice Clip, Participant Unaided Skill and Trial Number as main fixed effects, and Participant as random effect. We did not find any significant interaction effects. See Supplementary Table 1 for details. We report significant results at a stricter significance level (p<.005) to account for multiple comparisons.

Regarding emotion labeling in the Pre-AI Practice Trials, participants recognized some emotions better than others (Fig. 6a) and perceived different cues with varying accuracies (Fig. 6b), indicating that speaking rate and shrillness could be most verifiable in explanations, while pause proportion and pitch variation may be least. Furthermore, there was a wide range of average correctness among participants (M=49.9%, SD=18.3%), so we divided participants by whether they had above- or below-average unaided skill.

Analyzing the Main Trials, we found varying performances and perceptions due to different XAI Types (Fig. 7). Although participants may select a wrong emotion label as most likely, they may still select the correct label with low confidence in the balls and...
were rated as more helpful than None, though Saliency was only marginally, \( p = .0160 \). Providing Contrastive Saliency explanations were most effective and significantly better than not providing any explanation (None), \( p = .0007 \). Omitting the cues (C.Sample) led to a decrease in decision quality such that the difference from None was marginal, \( p = .0160 \). Providing Contrastive Saliency explanations did not help to improve decision quality, and, surprisingly, neither did providing all explanations combined together. All XAI Types were rated as more helpful than None, though Saliency was only marginally so (\( p = .0443 \)). All participants were equally confident (\( p = \text{n.s.} \)) in their emotion labels across XAI Types (Median=2 on -3 to 3 Likert scale). There was no difference in task time to label the emotions, though the most complex explanations (Saliency + C.Sample + Cues) was only 4.1 sec longer than None (26.1 vs. 21.0s).

### 5.3.4 Qualitative Results

We report why participants found specific XAI Types helpful or unhelpful and how they used them. Some participants depended on their own ability than rely on explanations, e.g., “I don’t think that the [Saliency] explanation information is helpful. I think that the voice is all you really need to be able to determine an emotion.” [P169], “I didn’t [use C.Sample] for the most part. I trust my own instincts.” [P54], “I think [Saliency + C.Sample + Cues] took longer than just listening to the clip ... I glance over it, but it doesn’t affect my decision as much” [P121].

Some participants struggled to use the Contrastive Saliency map. P26 found it “difficult to parse ... hard to analyze it by eye”. Errors in the Saliency explanations also led to distrust, as described by P8 that “the highlighted moments for Fearful don’t match well with [the] voice”. Conversely, the sophistication of the explanation led to over-trusting, with P146 mentioning that it was “helpful to view the color bar to determine which part has the most importance”, yet, this shallow interpretation led to him labeling wrongly. P117 commented that she was “unable to listen to different ratings the system has given to each emotion”, indicating her desire to hear other samples.

Counterfactual Sample explanations were more appreciated and marginally effective in improving decision quality. P38 felt that the “emotion in the clip is very clearly anger and it helped to hear the system show me what this voice would sound like when angry”; thus, she was matching samples by their perceived similarity. Similarly, P132 “first made my own judgment to narrow down the possible emotions, then listen to those emotions. I rate the one that matches the highest.” In contrast, P14 felt that C.Sample was “helpful to tell the difference between the neutral and calm voice” and “tried to see if there was a change in inflection or speed”. P103 felt that “it is slightly far away from the sample clip, every single one of them”, suggesting that he would appreciate Counterfactual Synthetics which would be generated to be more similar. Finally, P54 demurred that “the explanation information doesn’t elaborate at all why it’s giving that determination, so it’s mostly not helpful”; this indicates the need for deeper semantic explanations which C.Sample + Cues provides.

Instead of manually perceiving similarities or differences in voice clips, participants could read the cue differences in the Counterfactual Sample + Cues explanation. Their analytical understanding improved, as demonstrated in the vocabulary of their rationalization; e.g., P119 had a “better sense of the speaker’s pitch, loudness”. The semantic knowledge provided by cues also helped to reduce cognitive burden, e.g., P168 “used the information to confirm something I feel ambiguous about or just to make a guess and not have to spend so much effort deciding between guesses.” Specifically, cues helped to focus participants’ analyses, e.g., P90 found the explanation “helpful in letting you figure out what qualities to try to isolate in the voice clip to decide on where it learns in terms of emotion.”

Finally, although participants perceived Saliency + Counterfactual Sample + Cues as helpful, it did not improve decision quality. Participants rationalized the explanations by describing its various components separately, e.g., “it helps pinpoint what parts to listen to” [P31, Saliency], “the sample clips for each emotion are [helpful]” [P163, C.Sample], “compare the voices and the levels (like shrillness and pitch)” [P15, Cues]. However, no one explicitly described multiple components together, and there were few explicit descriptions about the saliency map. Perhaps, participants could not focus on specific explanation details. P167 was “not sure how to apply cross the broad”, suggesting an issue with information overload.

### 5.4 Summary of Results

We summarize the results from our three evaluation studies. The modeling study showed that RexNet provides relevant Saliency explanations, accurate Counterfactual Cues explanations, and promising Counterfactual Synthetics. These explanations helped to improve RexNet’s performance over the base CNN. The think-aloud user study showed that RexNet explanations align with how users innately perceive and infer vocal emotions, validating the XAI Perceptual Processing framework. We identified limitations in user
We chose to evaluate with vocal emotion recognition since it is an everyday task that is feasible to test with lay users. However, most people are already innately skilled in this, so this diminishes their need for AI or XAI to help them. Conversely, relatable explanations need to be relatable and contextualized to be more meaningfully interpreted. Specifically, we supported four criteria for relatability: contrastive concepts, saliency, counterfactuals, and associated cues.

To address the weaknesses of some relatable explanations, we discuss ways to further improve their effectiveness. Using counterfactual syntheses, instead of counterfactual samples would refine the difference between the example and target, so this may focus the user’s attention to more meaningful differences and improve discriminating between concepts. Moreover, our current approach identifies one set of cue differences across multiple salient locations. Instead, different cue sets can be associated with specific highlights in the saliency map. This can provide more semantics to various parts of a saliency map, to indicate why particular regions were important, and improve the usefulness of saliency maps.

We chose to evaluate with vocal emotion recognition since it is an everyday task that is feasible to test with lay users. However, most people are already innately skilled in this, so this diminishes their need for AI or XAI to help them. Conversely, relatable explanations may be more useful for more analytical tasks and applications with more explicit domain knowledge (e.g., engine noise diagnosis). We had identified several potential confounds — fatigue, skill at recognizing emotions, participants copying system predictions, learning effects from exposure to prior XAI versions — and discuss how we mitigated them. 1) We controlled for fatigue by: a) providing breaks between sessions, b) randomizing instances across trial numbers. We checked for fatigue by measuring: a) repeatedly identical responses (no participants were disqualified), b) decreases in labeling correctness over trials (no significant difference). 2) We controlled for recognition skills by measuring labeling performance without XAI (Pre-AI) and analyzed our results with that as a factor. 3) A more realistic use of AI is for it to make predictions and the user would verify its decision. However, in a pilot study evaluating with this task, we found that participants may copy the prediction rather than study the explanation, thus leading to over-trusting [110] and diminishing the usefulness of explanations to improve decision quality. We mitigated copying by evaluating with a human-simulatability task, instead of a predicted label verification task, though this trades-off some ecological validity. 4) We mitigated learning effects by designing the experiment as between-subjects, otherwise, participants may exploit new knowledge in subsequent experiment conditions (with weaker explanations).

This work is the first to explore relatable explanations for vocal emotion prediction, with an initial set of cues and adequate explanation accuracy. Future work can leverage other vocal stimulus types and prosodic attributes [53], such as non-verbal expressions, affect bursts, and lexical information. In particular, we learned that participants focus on the change in tone in voices to infer emotion, so this should be included as a vocal cue. Counterfactual synthesis accuracy can be improved by using newer generators, such as Sequence-to-Sequence Voice Conversion [42], StarGAN-VC v2 [43]. Though generated from a unified architecture, the explanations still had some inconsistencies. Annotating and debiasing explanations [57, 113] could help to align explanations with user expectations [93] and improve the coherence between explanation types. Contrastive Cue relations were encoded as a table, but they could be represented as another data structure (e.g., decision trees or causal graphs) to better fit human mental models. Finally, further testing could evaluate the usage and usefulness of predictions and explanations in in-the-wild applications [63], such as with smart speakers (e.g., Amazon Echo) [70], smartphone digital assistants for mental health or emotion monitoring [9, 106], or AI coaching for call center employees [28, 68, 80].

Although many XAI techniques have been recently developed, many remain too technical, or focus on supporting data scientists and machine learning model developers. Instead, there is a growing call to support different stakeholders and less technical users [14, 21, 58] Towards this end, we have studied human perception and cognition to determine new requirements for XAI. Miller had argued for contrastive and counterfactual explanations based on philosophical and psychological principles [72]. This was extended by Wang et al. to identify human reasoning pathways that can be supported by specific XAI techniques [107]. We extend these perspectives by identifying a broader requirement that explanations need to be relatable and contextualized to be more meaningfully interpreted. Specifically, we supported four criteria for relatability: contrastive concepts, saliency, counterfactuals, and associated cues. Extending our work, explanations can be made more relatable by providing for other criteria such as: social proof [17, 72], narrative
stories [96] or rationalizations [21], analogies [24], user-defined concepts [25, 47, 116], and plausible explanations [93]. Moreover, human cognition has natural flaws, like cognitive biases and limited working memory. XAI should include designs and capabilities to mitigate cognitive biases [104], moderate cognitive load [2], and accommodate information handling preferences [105]. Relatable explanations may need to account for these human factors to communicate why they may deviate from human reasoning.

The XAI Perceptual Processing Framework was inspired by human perceptual reasoning, rather than higher-level cognition. The latter is relevant for complex decision-making tasks, such as doctors’ reasoning with disease models, which are specific cases of causal structural models. Wang et al. proposed the XAI Reasoning Framework based on human reasoning processes [104], but this was not explicitly implemented in a single machine learning architecture. The Intelligibility Toolkit [60] provided an API to automatically generate explanations to a taxonomy of questions [58, 59], but this was not implemented for deep learning. Future work can explore a meta-model that combines perceptual and reasoning faculties for more complex, human-like model explanations.

6.5 Generalization of Relatable XAI

Although we implemented RexNet for the application of vocal emotion recognition, the XAI Perceptual Processing Framework is generalizable to other audio and visual prediction applications. Other audio applications include equipment monitoring via vibrations [108], and heart murmur diagnosis [89], but this was not implemented for deep learning. Future work can explore a meta-model that combines perceptual and reasoning faculties for more complex, human-like model explanations.

7 CONCLUSION

We presented the XAI Perceptual Processing Framework to unify a set of contrastive, saliency, counterfactual and cues explanations towards relatable explainable AI. The framework was implemented with RexNet, a modular multi-task deep neural network with multiple explanations, trained to predict vocal emotions. From qualitative think-aloud and quantitative controlled studies, we found varying usage and usefulness across the relatable contrastive explanations. This work gives insights into providing and evaluating relatable contrastive explainable AI for perception applications, and contributes a new basis towards human-centered XAI.

ACKNOWLEDGMENTS

This work was supported in part by the Ministry of Education, Singapore under the grant T2EP20121-0040, and was carried out at the NUS Centre for Research in Privacy Technologies (N-CRIPT) and the NUS Institute for Health Innovation and Technology (iHealthtech).

REFERENCES

[1] Ashraf Abdul, Jie Vermeulen, Danding Wang, Brian Y Lam, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible translation environment. In Proceedings of the 2018 CHI conference on human factors in computing systems. 1–18.

[2] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–14.

[3] Amina Adadi and Mohammed Berrada. 2018. Seeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access (2018), 52138–52160.

[4] Purvi Agrawal and Srimat Ganapathy. 2020. Interpretable representation learning for speech and audio signals based on relevance weighting. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020), 2823–2836.

[5] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhavi, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In International conference on machine learning. PMLR. 173–182.

[6] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souli Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. 2020. Mental models of mere mortals with explanations of reinforcement learning. ACM Transactions on Interactive Intelligent Systems (TiiS) 10, 2 (2020), 1–37.

[7] Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siam Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58 (2020), 82–115.

[8] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Daniel Nolte, Kris Luyten, Karin Coninx, Iulianna Van der Sluis, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the 2018 IEEE conference on computer vision and pattern recognition. 8789–8797.

[9] Robert B Cialdini, Wilhelmina Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–14.

[10] Edward C. Carterette and Morton P. Friedman (Eds.). 1978. The rational imagination: How people create alternatives to reality. MIT press.

[11] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In Proceedings of the 24th international conference on intelligent user interfaces. 258–262.

[12] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernando Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–14.

[13] Edward C. Carterette and Morton P. Friedman (Eds.). 1978. The rational imagination: How people create alternatives to reality. MIT press.

[14] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O’Connell, Terrance Gray, Jianlin Cheng, and Gianluca Pollastri. 2008. A neural network for speech and audio signals based on relevance weighting. IEEE/ACM Transactions on Audio, Speech, and Language Processing 16, 10 (2008), 2580–2591.

[15] Paul Chun, Moonsu Kim, Joon Ho Lee, and Sang-Hyuk Shim. 2008. Some additional results for the n-gram model. In Proceedings of the 2008 IEEE international joint conference on natural language processing. 1279–1284.

[16] Junyong Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 8789–8797.

[17] Robert B Cialdini, Wilhelmina Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–14.

[18] Sven Coppers, Jan Van den Bergh, Kris Luyster, Karin Coninx, Iulianna Van der Sluis, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. Intellingly: an intelligible translation environment. In Proceedings of the 2018 CHI Conference
Towards Relatable Explainable AI with the Perceptual Process

CHI ’22, April 29-May 5, 2022, New Orleans, LA, USA

on Human Factors in Computing Systems. 1–13.

T Edward Damer. 2012. Attacking faulty reasoning. Cengage Learning.

Amit Dhurandhar, Fur-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Kirthikayen Shanmugam, and Pavlo Das. (n.d.). Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. Ann Arbor 1001 (n.d.), 8109.

Upol Ehsan, Brent Harrington, Larry Chan, and Mark O Riedel. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 81–87.

Malin Eibhard, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In 23rd international conference on intelligent user interfaces. 211–223.

Dedre Gentner and Linsey Smith. 2012. Analogical reasoning. Encyclopedia of human behavior 2 (2012), 130–136.

Amirata Ghobiani, James Weider, James Y Zou, and Been Kim. 2019. Towards Automatic Concept-based Explanations. Advances in Neural Information Processing Systems 32 (2019), 9277–9286.

Daniel G Goldstein and David Rothschild. 2014. Lay understanding of probability distributions. Judgment & Decision Making 9, 1 (2014).

E Bruce Goldstein. 2014. Cognitive psychology: Connecting mind, research and everyday experience. Cengage Learning.

Cristina Gorrostueta, Richard Bruttì, Kye Taylor, Avi Shapiro, Joseph Moran, Ali Azarbayejani, and John Kane. 2018. Attention-based Sequence Classification for Affected Detection. In InterSpeech. 506–510.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In International Conference on Machine Learning. PMLR, 2376–2384.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51, 5 (2018), 1–42.

David Gunning and David Aha. 2019. DARPA’s explainable artificial intelligence (XAI) program. AI Magazine 40, 2 (2019), 44–58.

Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2014. Speech version. IEEE/ACM Transactions on Audio, Speech, and Language Processing 115 (2018), 101–106.

Lisa Anne Hendrickz, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating Counterfactual Explanations with Natural Language. In ICML Workshop on Human Interpretability in Machine Learning. 95–98.

Shawn Hershley, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51, 5 (2018), 1–42.

Denis J Hilton. 1990. Conversational processes and causal explanation. Psychological Bulletin 107, 1 (1990), 65.

Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual analytics in deep learning: An interrogative survey for the next frontiers. IEEE transactions on visualization and computer graphics 107, 1 (1990), 65.

Neha Jain, Shishir Kumar, Amit Kumar, Pourya Shamsolmoali, and Masoumeh Zareapoor. 2018. Hybrid deep neural networks for face emotion recognition. Pattern Recognition Letters 115 (2018), 101–106.

Patrick J Juslin and Petri Laukka. 2001. Impact of intended emotion intensity on everyday experience. Cognitive psychology: Connecting mind, research and everyday experience. Cengage Learning.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT. 4171–4186.

Brian Kim, Rajiv Khanna, and Ohlsswanam O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. Advances in neural information processing systems 29 (2016).

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In International Conference on Machine Learning. PMLR, 1858–1894.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25 (2012), 1097–1105.

Andreas Krug, René Knaebel, and Sebastian Stober. 2018. Neuron activation profiles for interpreting convolutional speech recognition models. In NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language (RASL).

Andreas Krug and Sebastian Stober. 2018. Introduction for convolutional automatic speech recognition. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. 187–199.

Adi Lausen and Kurt Hammerschmidt. 2020. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. Human computer interactions and Social Sciences Communications 7, 1 (2020), 1–17.

Thi Ha, Suhang Wang, and Dongwon Lee. 2020. GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model’s Prediction. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 238–248.

Benjamin Latham, Cynthia Rudin, Tyler H McCormick, and David Madigan. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics 9, 3 (2015), 1353–1371.

Chung-Yi Li, Pei-Chieh Yuan, and Hung-Yi Lee. 2020. What does a network layer hear? analyzing hidden representations of end-to-end asr through speech synthesis. In ICAASP-2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 6434–6438.

Kunpeng Li, Ziyan Wu, Xuan-Chuan Peng, Jan Ernst, and Yun Fu. 2018. Tell me where to look: Guided attention inference network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 9215–9223.

Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–15.

Brian Y Lim and Anind K Dey. 2009. Assessing demand for intelligibility in context-aware applications. In Proceedings of the 11th international conference on Ubiquitous computing. 195–204.

Brian Y Lim and Anind K Dey. 2010. Toolkit to support intelligibility in context-aware applications. In Proceedings of the 12th ACM international conference on Ubiquitous computing. 11–22.

Brian Y Lim and Anind K Dey. 2011. Design of an intelligible mobile context-aware application. In Proceedings of the 13th international conference on human computer interaction with Mobile devices and services. 157–166.

Brian Y Lim and Anind K Dey. 2021. Investigating intelligibility for uncertain context-aware applications. In Proceedings of the 13th international conference on Ubiquitous computing. 415–424.

Brian Y Lim and Anind K Dey. 2013. Evaluating intelligibility usage and usefulness in a context-aware application. In International Conference on Human-Computer Interaction. Springer. 92–101.

Brian Y Lim, Anind K Dey, and Daniel Arvrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 238–248.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4401–4410.

Xueming Luo, Marco Shaoojin Qin, Zheng Fang, and Zhe Qu. 2021. Artificial Intelligence Coaches for Sales Agents: Caveats and Solutions. Journal of Marketing 85, 2 (2021), 14–32.
A APPENDIX

A.1 Vocal cues for different emotions

Supplementary Fig. 1. Distribution of cue values for different emotions and the average across all voice clips. Values calculated from the RAVDESS dataset [66]. Differences were used to calculate cue difference relations. Grey line indicates average value.
A.2 User Study Survey

Supplementary Fig. 2. Tutorial to clarify users’ tasks, to interpret the “balls and bins” question [26] and screening question to check users’ audio equipment.
Supplementary Fig. 3. Tutorial on the contrastive saliency explanation and screening question to check users’ interpretation.

Supplementary Fig. 4. Tutorial on the counterfactual sample explanation.
Supplementary Fig. 5. Tutorial on the contrastive cue explanation and screening question to check users’ understanding about vocal cues.
Supplementary Fig. 6. Example practice session per-voice trial before revealing the system's XAI information (Pre-XAI).
Supplementary Fig. 7. Example main study per-voice trial without the system’s explanation.

Supplementary Fig. 8. Example main study per-voice trial with the contrastive saliency explanation.

Supplementary Fig. 9. Example main study per-voice trial with the counterfactual sample explanation.
Supplementary Fig. 10. Example main study per-voice trial with the counterfactual sample and contrastive cue explanations.

Supplementary Fig. 11. Example main study per-voice trial with the contrastive saliency, counterfactual sample and contrastive cue explanations.
Supplementary Fig. 12. Example main study per-voice trial with the questionnaire after revealing the system’s XAI information (Post-XAI).
Supplementary Fig. 13. Example practice session per-voice trial to show the correct answer and review users’ choices.
### A.3 User Study Analysis: Statistical Model

Supplementary Table 1. Statistical analysis of responses due to effects (one per row), as linear mixed effects models with random effects, fixed effects, and their interaction effect. $F$ and $p$ values indicate ANOVA tests and $R^2$ indicate model goodness-of-fit.

| Response | Linear Effects Model (Participants as random effects) | $F$  | $p>F$ | $R^2$ |
|----------|-------------------------------------------------------|------|-------|-------|
| **Labeling Correctness** | XAI Type + 4.2 <.0030 .371 | | | |
| | Participant Unaided Skill + 50.0 <.0001 | | | |
| | Emotion 68.6 <.0001 | | | |
| | Voice Clip + 67.1 <.0001 | | | |
| | Trial Number + 0.6 n.s. | | | |
| | Confidence Rating + 19.7 <.0001 | | | |
| | Helpfulness Rating 7.9 .0053 | | | |
| **Confidence on Correct Label** | XAI Type + 4.2 <.0029 .435 | | | |
| | Participant Unaided Skill + 64.4 <.0001 | | | |
| | Emotion 81.8 <.0001 | | | |
| | Voice Clip + 80.2 <.0001 | | | |
| | Trial Number + 0.3 n.s. | | | |
| | Confidence Rating + 51.9 <.0001 | | | |
| | Helpfulness Rating 6.4 .0113 | | | |
| **Confidence Rating** | XAI Type + 0.5 n.s. | | | |
| | Participant Unaided Skill + 0.1 n.s. | | | |
| | Emotion 4.0 .0002 | | | |
| | Voice Clip + 7.7 <.0001 | | | |
| | Trial Number 0.3 n.s. | | | |
| **Helpfulness Rating** | XAI Type + 5.7 .0002 .831 | | | |
| | Participant Unaided Skill + 6.6 .0114 | | | |
| | Emotion 2.2 .0283 | | | |
| | Voice Clip + 7.7 <.0001 | | | |
| | Trial Number 0.7 n.s. | | | |
| **Log_{10}(Task Time)** | XAI Type + 1.0 n.s. | | | |
| | Participant Unaided Skill + 2.8 n.s. | | | |
| | Emotion + 2.8 n.s. | | | |
| | Voice Clip + 9.3 <.0001 | | | |
| | Trial Number 4.5 .0001 | | | |