

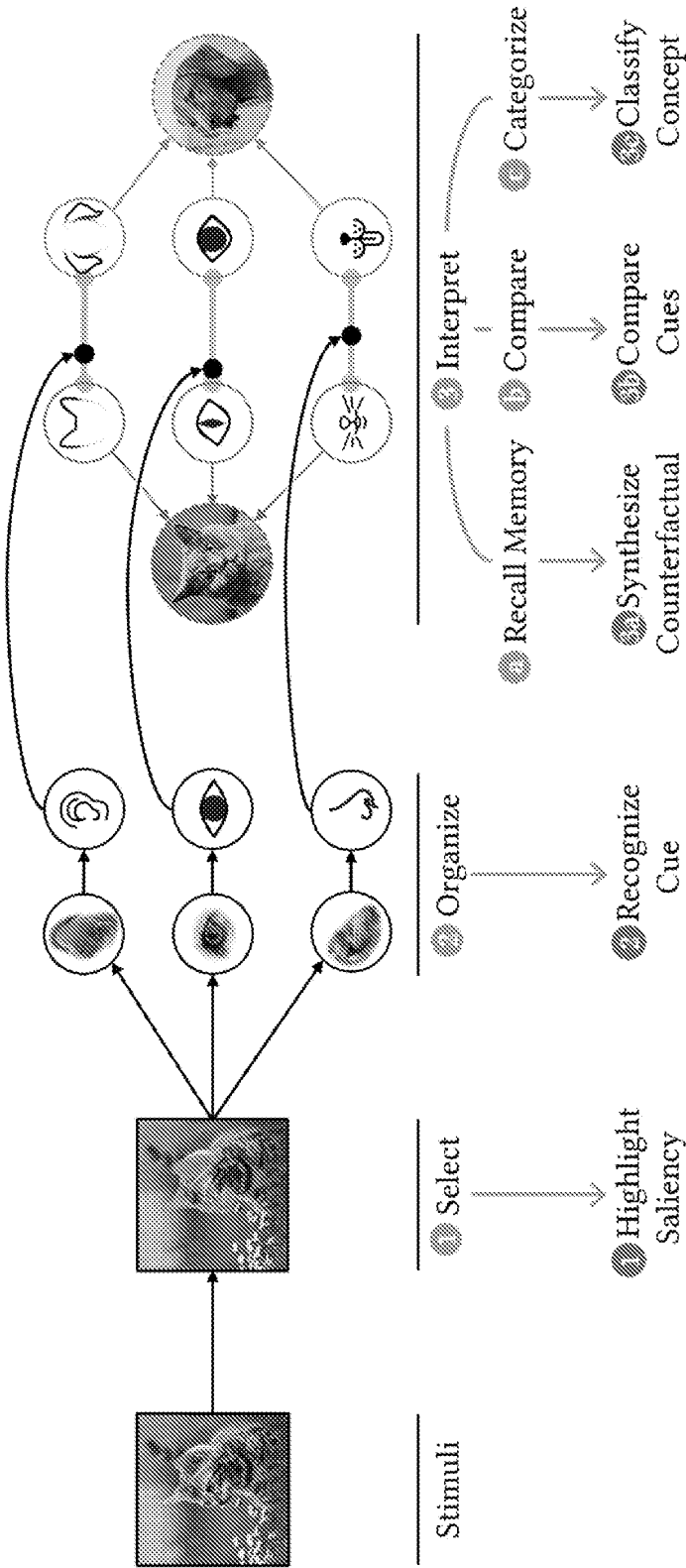


FIG. 1A

Vocal cues to characterize voices for emotion recognition.

Vocal Cue [39]	User Friendly 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}

FIG. 1B

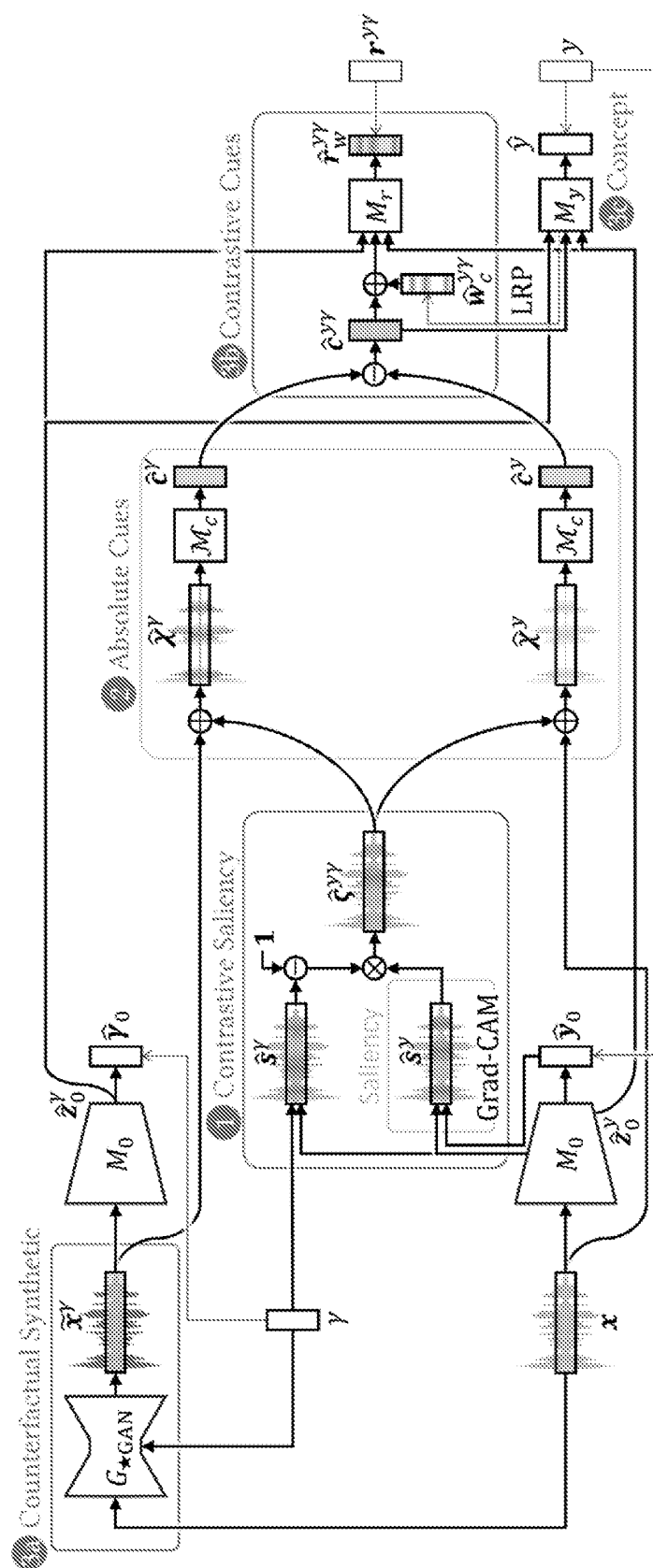
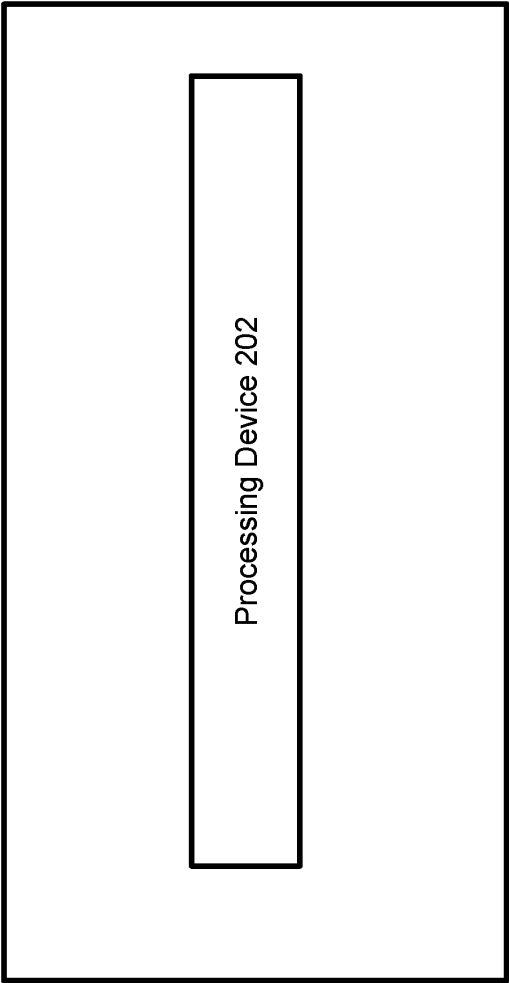


FIG. 2A



200

FIG. 2B

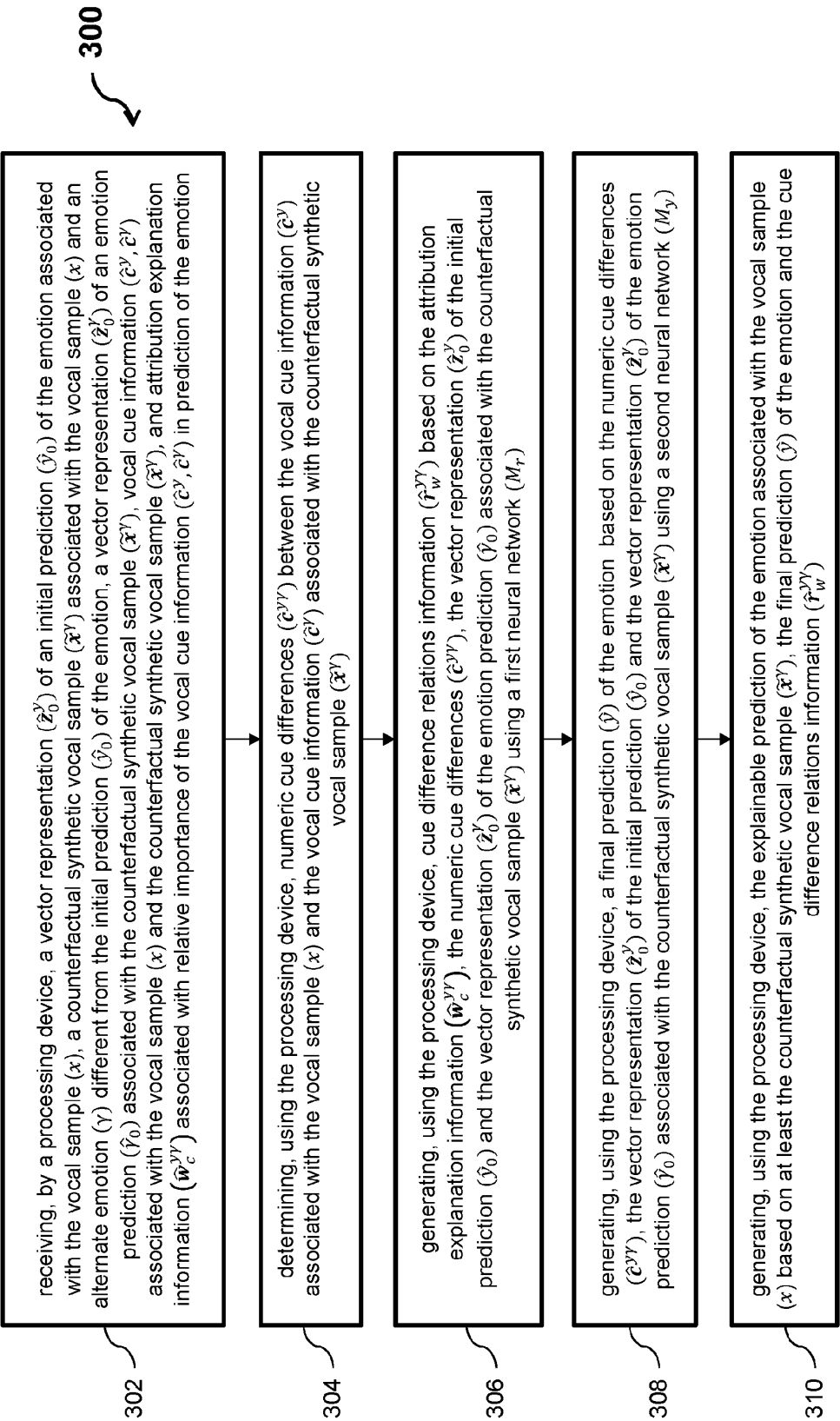


FIG. 2C

400

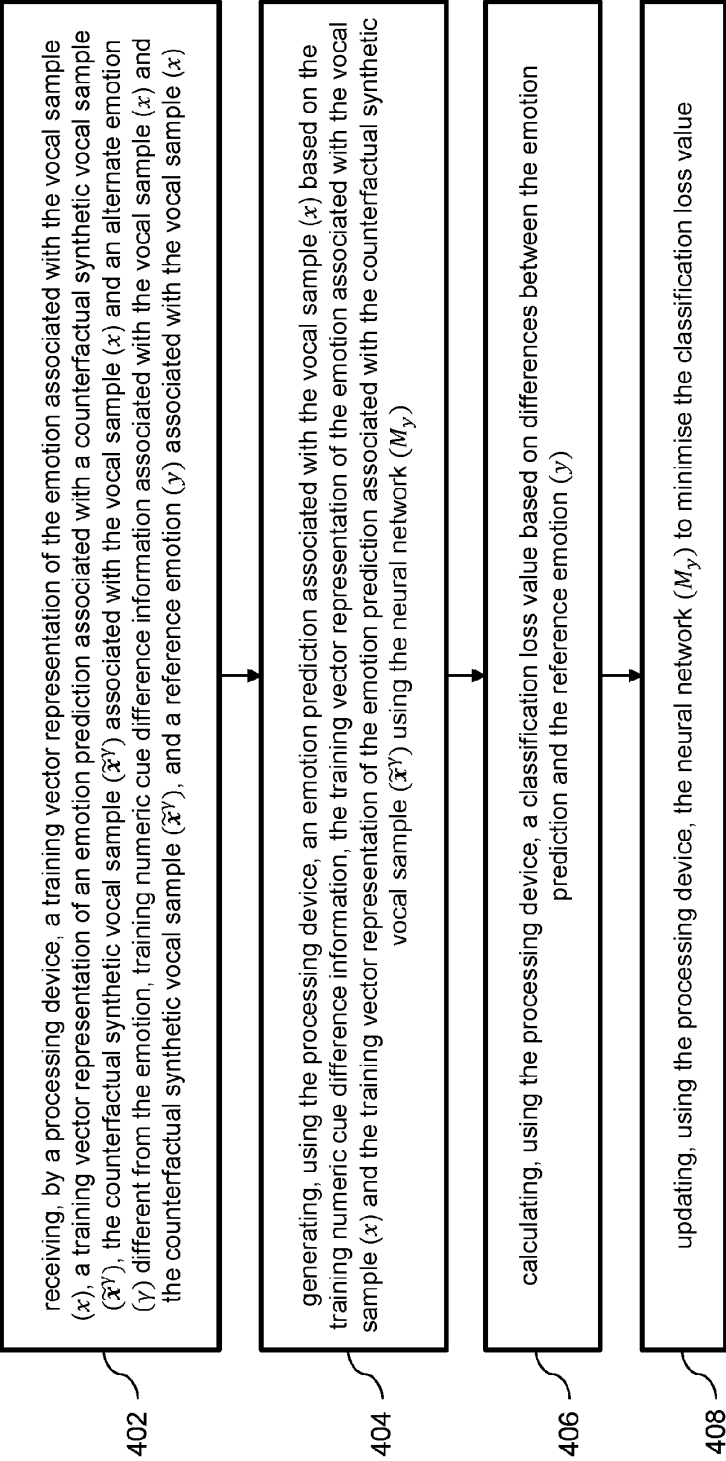


FIG. 2D

500

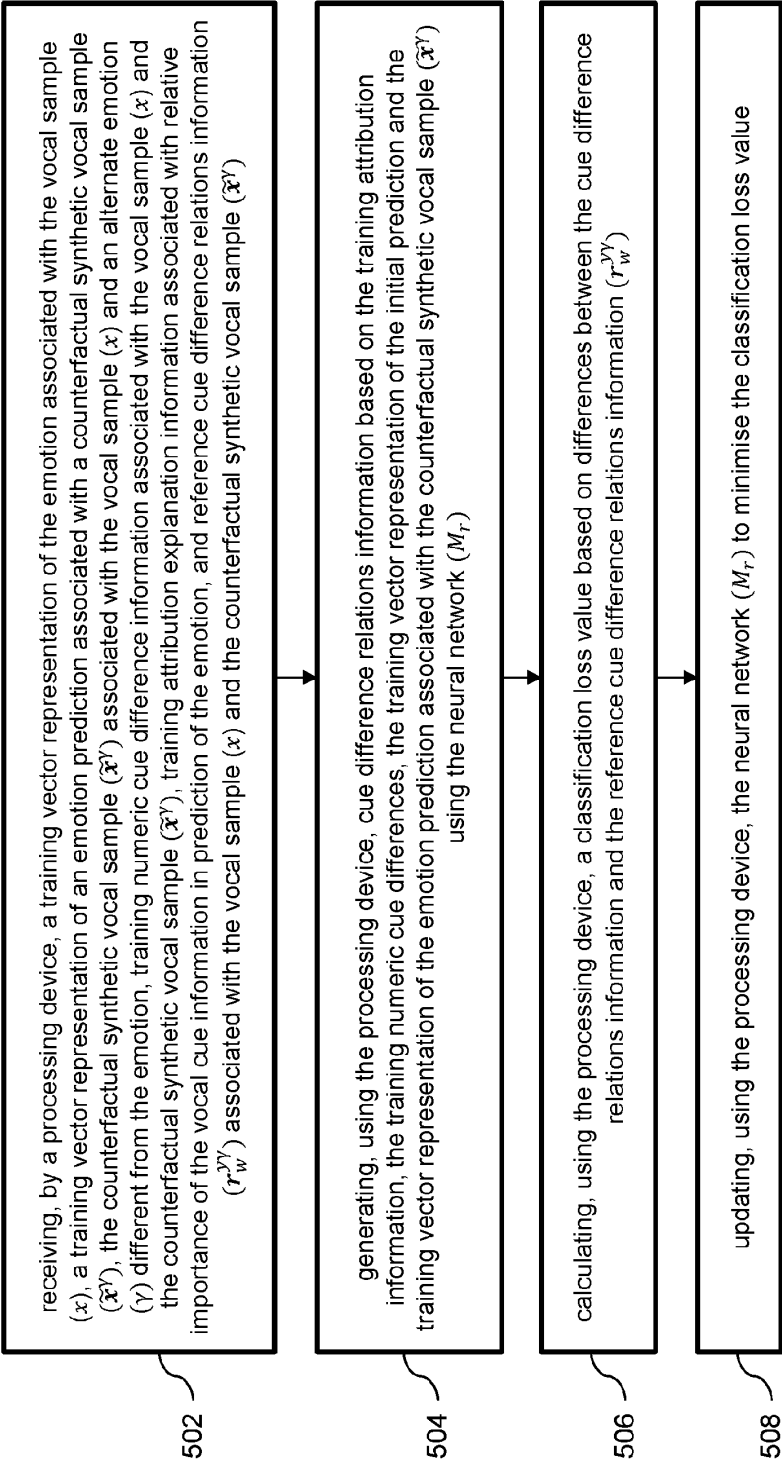


FIG. 2E

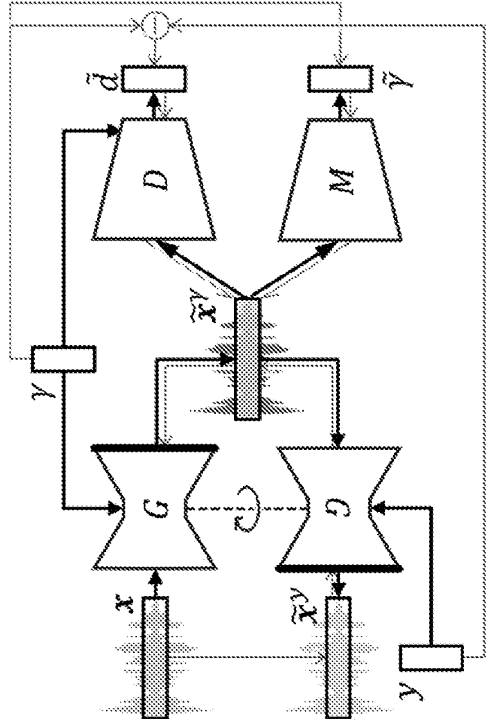


FIG. 3

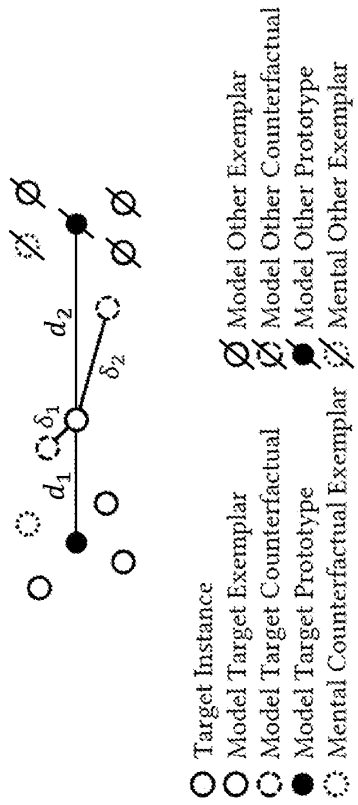


FIG. 4A

Vocal cues for each emotion relative to average levels. Values indicate the absolute cues.

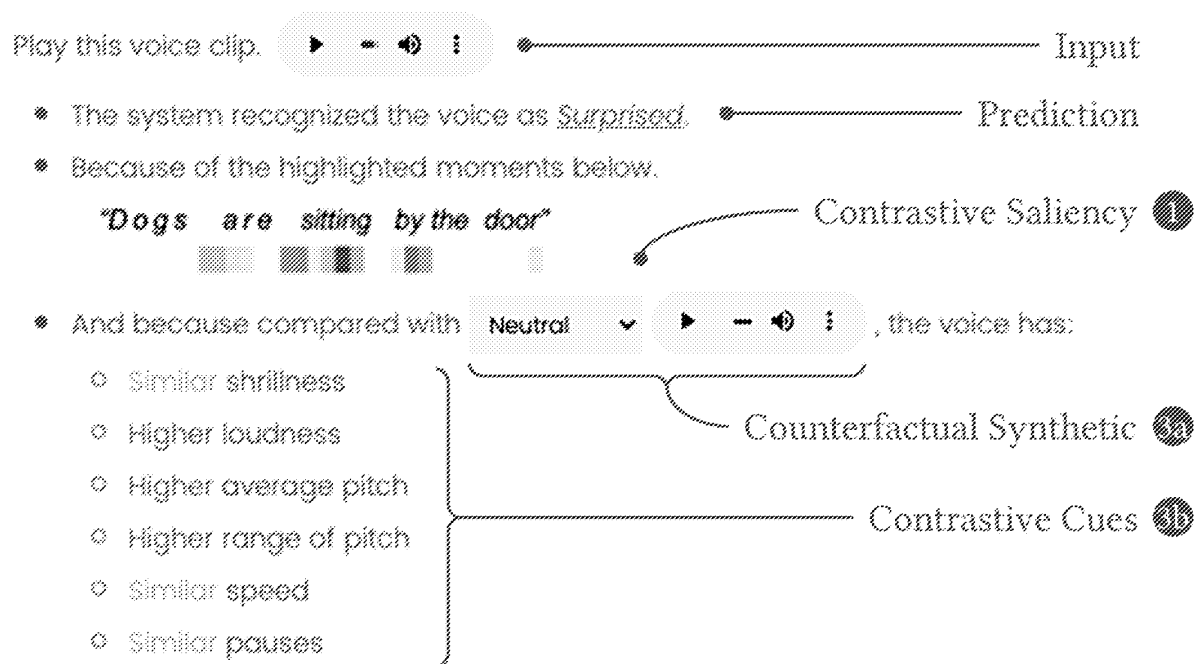
Target Emotion	Vocal Cue					
	Shrillness	Loudness	Average Pitch	Pitch Range	Speaking Rate	Proportion of Pauses
Neutral	Average	Low	Low	Low	High	Low
Calm	Low	Low	Low	Low	Average	Average
Happy	High	High	High	Average	Average	Low
Fearful	Average	High	High	High	High	Average
Surprised	Average	Average	Average	High	High	Low
Sad	Low	Low	Average	Average	Average	High
Disgust	Average	Average	Low	Low	Low	High
Angry	High	High	High	High	Low	Average

FIG. 4B

Vocal cues for target emotions compared to another emotion (Happy). Values indicate the contrastive cues.

Target Emotion	Vocal Cue						Compared to Emotion
	Shrillness	Loudness	Average Pitch	Pitch Range	Speaking Rate	Proportion of Pauses	
Neutral	Lower	Lower	Lower	Lower	Similar	Similar	Happy
Calm	Lower	Lower	Lower	Lower	Lower	Similar	Happy
Happy	Similar	Similar	Similar	Similar	Similar	Similar	Happy
Fearful	Similar	Similar	Similar	Similar	Similar	Higher	Happy
Surprised	Lower	Similar	Similar	Similar	Similar	Similar	Happy
Sad	Lower	Lower	Lower	Lower	Similar	Higher	Happy
Disgust	Average	Lower	Lower	Lower	Lower	Higher	Happy
Angry	Higher	Higher	Similar	Similar	Lower	Higher	Happy

FIG. 4C



- And because compared with **Neutral** , the voice has:
 - Similar shrillness
 - Higher loudness
 - Higher average pitch
 - Higher range of pitch
 - Similar speed
 - Similar pauses
- Counterfactual Synthetic 3a
- Contrastive Cues 3b

FIG. 5A

Evaluation results of model prediction performance and explanation correctness for RexNet and baseline models (Random, Base CNN). RexNet models compared include the full model, and one trained with Counterfactual Samples (C.Samples) without StarGAN used in user studies. Grey numbers calculated by definition, instead of from empirical results. * same as base CNN model.

Variable	Metric	Model			
		Random	Base CNN	RexNet	RexNet w/ C.Samples
Initial Concept $\hat{y}_0$	Emotion accuracy (8 classes)	12.5%	75.7%	79.5%	78.8%
Final Concept $\hat{y}$	Emotion accuracy (8 classes)			78.5%	77.4%
Absolute Saliency $\hat{s}^y$	Ablated accuracy decrease			14.9%	16.7%
Contrastive Saliency $\hat{s}^{yy}$	Ablated accuracy decrease			13.7%	16.3%
Counterfactual Synthetic $\hat{x}^y$	Reconstruction similarity			0.553	1
	Identity accuracy (24 classes)	4.2%		60.2%	96.2%
	Emotion accuracy (8 classes)	12.5%		30.7%	78.7%
Cue Difference Relation $\hat{F}_w^{yy}$	Cue accuracy (3 classes, 6 labels)			71.9%	71.6%

FIG. 5B

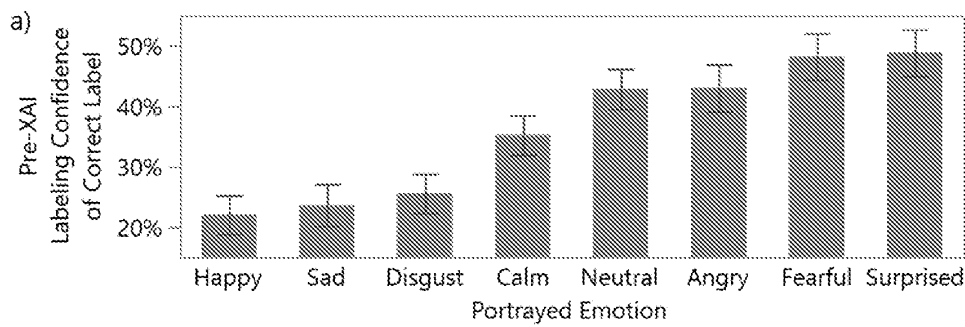


FIG. 6A

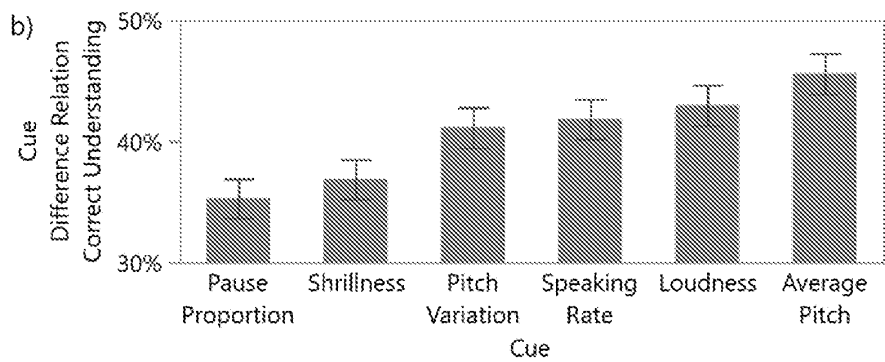
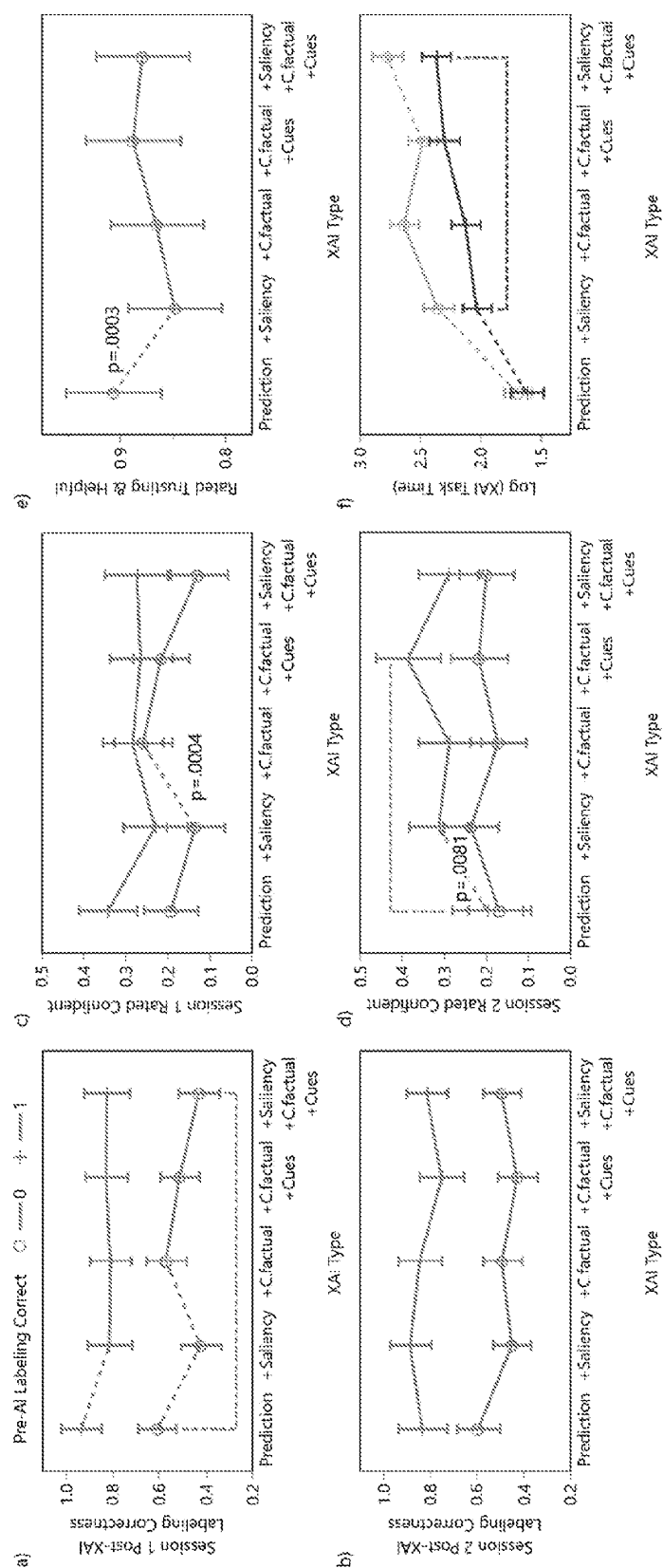
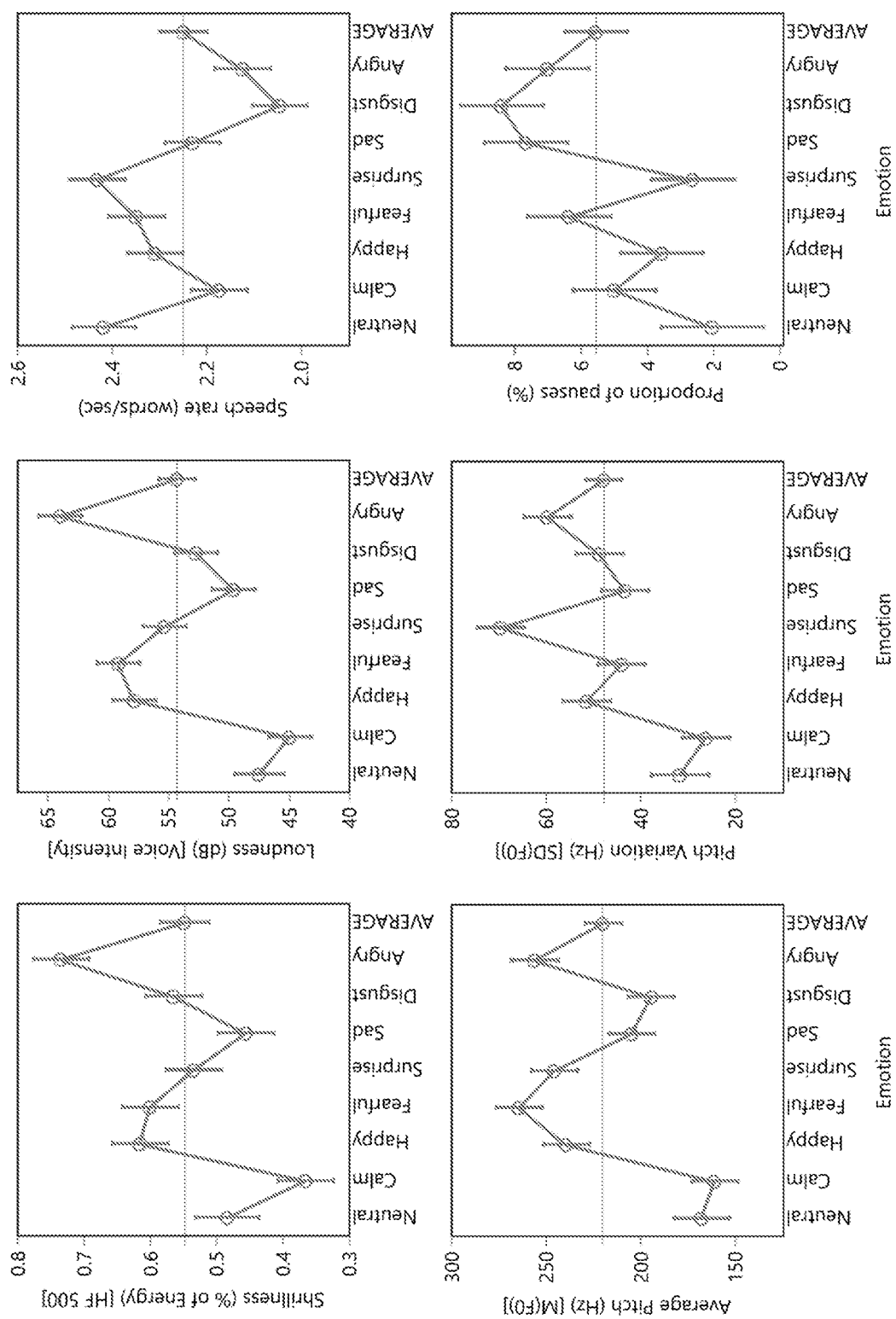


FIG. 6B



FIGS. 7A-7F



Choice with Sliders

Sometimes it is unclear which is the dominant emotion, and you may think several emotions are likely correct. We use a set of sliders to indicate which emotion is likely to be the dominant emotion. Suppose you think that the speech is 70% likely Neutral, 20% likely Angry, and 10% likely Surprised, then you will indicate the sliders as shown below.

Note that % likelihoods need to sum to 100.

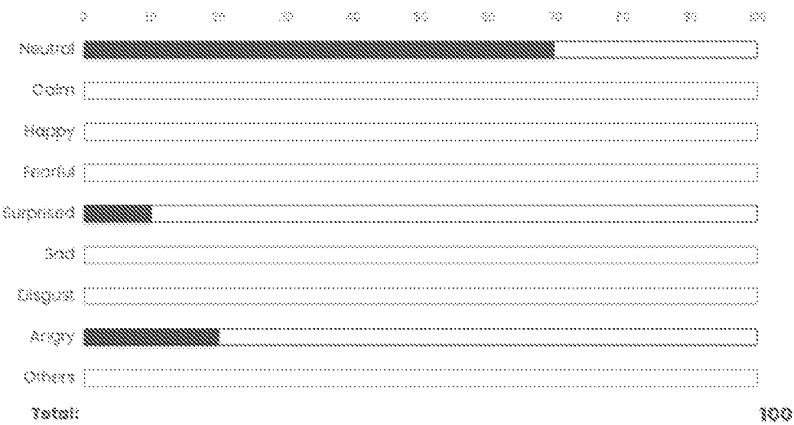
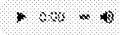


FIG. 9

Tutorial: Smart System

Given the emotional voice clip: 

A Smart System is developed to recognize emotional states from speech. The system firstly indicates its predicted emotion category.

- The system recognized the voice as Sad.

Note that the system's prediction may be inaccurate sometimes, it's up to you to trust it or not.

Q1. What does the person say in the previous clip?

- ☐ "Kids are talking by the door."
- ☐ "Kids are talking in the hall."
- ☐ "Kids are playing with the ball."

FIG. 10

Importance Indicator

The system uses an **importance indicator** to highlight why it predicts a certain emotion category or why it predicts one category instead of another. The information is shown in a colored bar below the transcript. Darker  means the system thinks it's more important.

- The system recognized the voice as Sad instead of Neutral ▾.
- Because of the highlighted moments below.
 "Kids are talking by the door"
-  

You may want to pay attention on highlighted moments when making your decision.

Q2. In the example above, which word is most important when the system predicts the voice as Sad instead of Surprised?

door

☐

talking

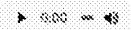
☐

the

☐

FIG. 11

Comparison Voice Clip

When judging the target voice clip  , you may use **comparison voice clips** for reference. It has the same words but sounds in a different emotion style, e.g.

- The system recognized the target voice as Sad.
- The voice sounds like  if it was Neutral ▾.

When making decision, you may compare examples from different emotion categories.

FIG. 12



Voice Cues

To judge the target voice clip , the system generates descriptive **Voice Cues**.

Voice Cues	Description
Average pitch	Pitch refers to the degree of highness of the tone.
Range of pitch	
Loudness	Loudness of the speech.
Shrillness	Percentage of high frequency energy of the voice.
Speed	Speed of the speech.
Pause	Proportion of pauses.

Each cue can be described as *lower (fewer) / similar / higher (more)* by comparing the target voice with a voice in another emotion category, e.g.

- The system recognized the target voice as Sad.
- Because compared with Neutral , the voice has:
 - Similar shrillness
 - Similar loudness
 - Similar average pitch
 - Similar range of pitch
 - Lower speed
 - More pauses

When making decision, you may compare the description about cues in different emotion categories.

Q3. Which voice has higher / more ?

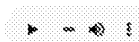
	Voice 1 	Voice 2 
Shrillness	☐	☐
Loudness	☐	☐
Average pitch	☐	☐
Range of pitch	☐	☐
Speed	☐	☐
Pauses	☐	☐



FIG. 13

Voice 1

Play this voice clip. (Dogs are sitting by the door)

Q1. Which dominant emotion category do you think it belongs to?
Note: at least one choice must be more than 0%, and all % likelihoods need to sum to 100.

	0	10	20	30	40	50	60	70	80	90	100
Neutral											
Calm											
Happy											
Fearful											
Surprised											
Sad											
Disgust											
Angry											
Others											
Total:											0

FIG. 14



Voice 1

Play this voice clip.

◆ The system recognised the voice as Surprised.

Please study the information carefully before moving to the questionnaire page.

FIG. 15



Voice 1

Play this voice clip. 

- * The system recognized the voice clip as *Surprised* instead of *Neutral*.
- * Because of the highlighted moments below.
"Dogs are sitting by the door"


☒ Neutral

☐ Calm

☐ Happy

☐ Fearful

☐ Sad

☐ Disgust

☐ Angry

Please study the information carefully before moving to the questionnaire page.

FIG. 16



Voice 1

Play this voice clip. 

- * The system recognized the voice as *Surprised*.
- * The voice sounds like  if it is

☒ Neutral

☐ Calm

☒ Surprised

☐ Fearful

☐ Surprised

☐ Sad

☐ Disgust

☐ Angry

Please study the information carefully before moving to the questionnaire page.

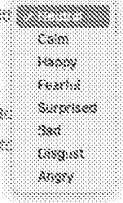
FIG. 17



Voice 1

Play this voice clip. 

• The system recognized the voice as Surprised.

• Because compared with  , the voice has:

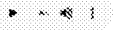
- Similar shrillness
- Higher loudness
- Higher average pitch
- Higher range of pitch
- Similar speed
- Similar pauses

Please study the information carefully before moving to the questionnaire page.

FIG. 18



Voice 1

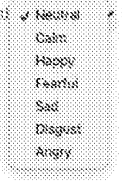
Play this voice clip. 

• The system recognized the voice as Surprised.

• Because of the highlighted moments below:

"Dogs are sitting by the door"



• And because compared with  , the voice has:

- Similar shrillness
- Higher loudness
- Higher average pitch
- Higher range of pitch
- Similar speed
- Similar pauses

Please study the information carefully before moving to the questionnaire page.

FIG. 19



Q2. Which dominant emotion do you think the voice  belongs to?

Note: at least one choice must be more than 0%, and all % likelihoods need to sum to 100.

010203040506070809100

Neutral

Calm

Happy

Fearful

Surprised

Sad

Disgust

Angry

Others

Total:

Q3. The voice  has ... (cue) ... that is ... (higher / lower / similar) ... than happy.

	Lower	Similar	Higher	Cannot tell
Steadiness	☐	☐	☐	☐
Lowness	☐	☐	☐	☐
Average pitch	☐	☐	☐	☐
Range of pitch	☐	☐	☐	☐
Speed	☐	☐	☐	☐
Pauses	☐	☐	☐	☐

Q4. Do you agree or disagree with the following statements?

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I am confident of my emotion choice.	☐	☐	☐	☐	☐	☐	☐
I trust the system's prediction.	☐	☐	☐	☐	☐	☐	☐
The system is helpful.	☐	☐	☐	☐	☐	☐	☐

FIG. 20



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 and Voice clips as random effects)	F	p>F	<i>R</i> ²
Post XAI Labling Correctness	XAI Type +	7.3	<.0001	.497
	Pre-XAI Labeling Correct +	503.4	<.0001	
	Pre-XAI Labeling Correct × XAI Type +	3.3	.0098	
	Trial Number +	1.4	n.s.	
	Session +	0.2	n.s.	
	XAI Type × Session +	2.4	.0442	
	Pre-XAI Labeling Correct × Session +	0.005	n.s.	
	Pre-XAI Labeling Correct × XAI Type × Session	2.0	n.s.	
Log (XAI Task Time)	XAI Type +	121.9	<.0001	.504
	Pre-XAI Labeling Correct +	19.1	<.0001	
	Pre-XAI Labeling Correct × XAI Type +	5.1	.0004	
	Trial Number +	35.2	<.0001	
	Session +	70.6	<.0001	
	XAI Type × Session +	6.1	<.0001	
	Pre-XAI Labeling Correct × Session +	1.2	n.s.	
	Pre-XAI Labeling Correct × XAI Type × Session	0.9	n.s.	
Rated Confident	XAI Type +	2.3	n.s.	.569
	Pre-XAI Labeling Correct +	62.5	<.0001	
	Pre-XAI Labeling Correct × XAI Type +	0.6	n.s.	
	Trial Number +	2.6	.0120	
	Session +	0.9	n.s.	
	XAI Type × Session +	6.2	<.0001	
	Pre-XAI Labeling Correct × Session +	0.05	n.s.	
	Pre-XAI Labeling Correct × XAI Type × Session	4.3	.0020	
Rated Trusting & Helpful	XAI Type +	4.0	.0032	.492
	Pre-XAI Labeling Correct +	69.5	<.0001	
	Pre-XAI Labeling Correct × XAI Type +	1.4	n.s.	
	Trial Number +	0.8	n.s.	
	Session +	0.7	n.s.	
	XAI Type × Session +	1.0	n.s.	
	Pre-XAI Labeling Correct × Session +	0.3	n.s.	
	Pre-XAI Labeling Correct × XAI Type × Session	0.2	n.s.	

FIG. 21

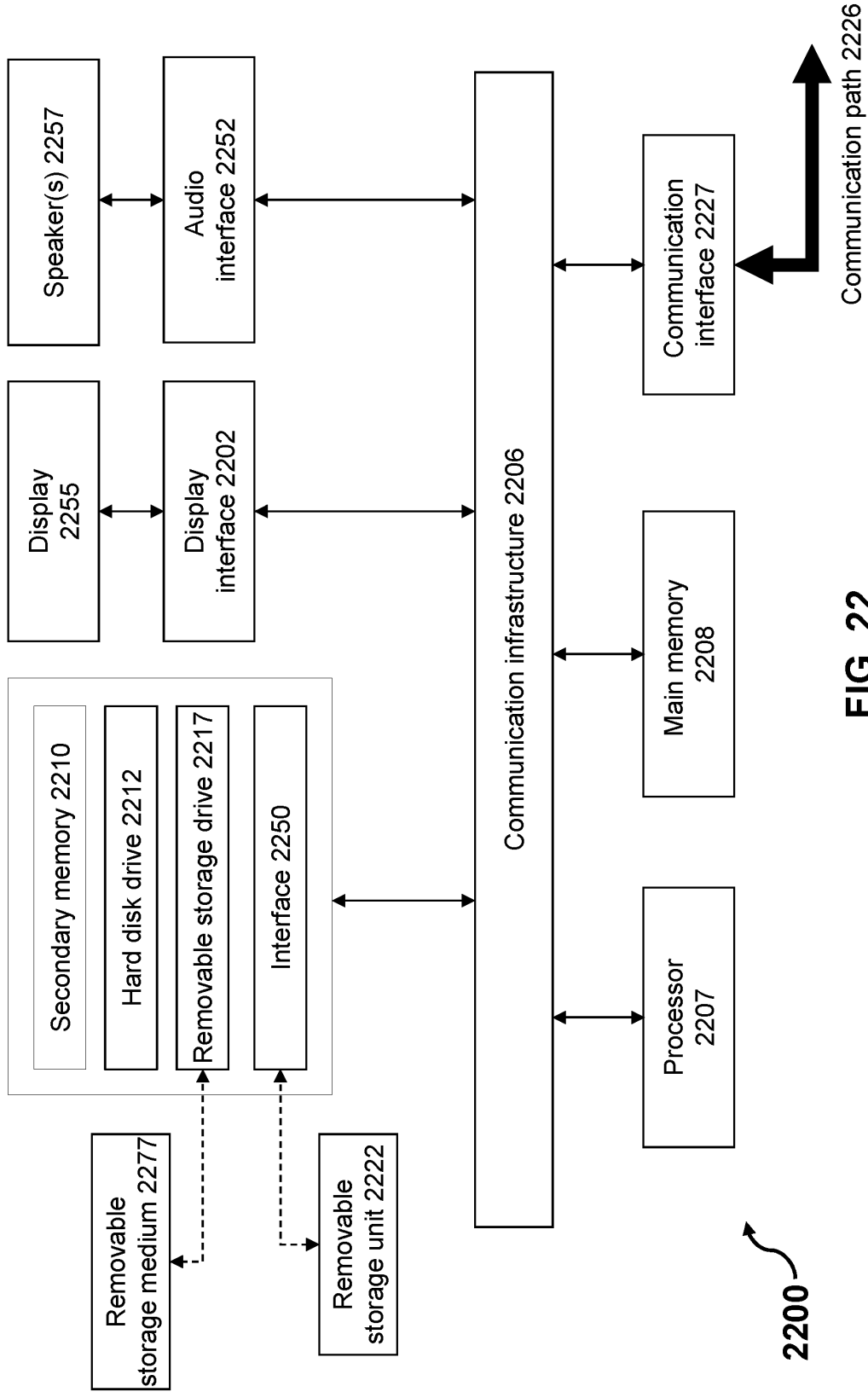


FIG. 22