synthetic vocal sample, and attribution explanation information associated with relative importance of the vocal cue information in prediction of the emotion. The method also includes determining, using the processing device, numeric cue differences between the vocal cue information associated with the vocal sample and the vocal cue information associated with the counterfactual synthetic vocal sample, generating, using the processing device, cue difference relations information based on the attribution explanation information, the numeric cue differences , the vector representation of the initial prediction and the vector representation of the emotion prediction associated with the counterfactual synthetic vocal sample using a first neural network, generating, using the processing device, a final prediction of the emotion based on the numeric cue differences, the vector representation of the initial prediction and the vector representation of the emotion prediction associated with the counterfactual synthetic vocal sample using a second neural network, and generating, using the processing device, the explainable prediction of the emotion associated with the vocal sample based on at least the counterfactual synthetic vocal sample, the final prediction of the emotion and the cue difference relations information.

[0005] The step of receiving the counterfactual synthetic vocal sample can include generating, using the processing device, the counterfactual synthetic vocal sample based on the vocal sample and the alternate emotion using a generative adversarial network.

[0006] The step of receiving the vocal cue information associated with the vocal sample and the counterfactual synthetic vocal sample can include generating, using the processing device, a contrastive saliency explanation based on the vocal sample, the initial prediction, and the alternate emotion using a visual explanation algorithm, and determining, using the processing device, the vocal cue information associated with the vocal sample based on the vocal sample and the contrastive saliency explanation, and the vocal cue information associated with counterfactual synthetic vocal sample based on the counterfactual synthetic vocal sample and the contrastive saliency explanation.

[0007] The vocal cue information can be associated with one or more of a group consisting of: shrillness, loudness, average pitch, pitch range, speaking rate and proportion of pauses.

[0007a] The cue difference relations information ($\hat{\boldsymbol{r}}_w^{\gamma\gamma}$) can include ordinal cue difference relations information ($\hat{\boldsymbol{r}}_w^{\gamma\gamma}$).

[0007b] The counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$), the cue difference relations information ($\hat{\boldsymbol{r}}_w^{\gamma\gamma}$), and the final prediction ($\hat{y}$) of the emotion can be generated in a multi-task, multi-modal architecture.

[0008] Another aspect of the present disclosure provides a system for generating an explainable prediction of an emotion associated with a vocal sample. The system can include a processing device configured to receive a vector representation of an initial prediction of the emotion associated with the vocal sample, a counterfactual synthetic vocal sample associated with the vocal sample and an alternate emotion different from the initial prediction of the emotion, a vector representation of an emotion prediction associated with the counterfactual

synthetic vocal sample, vocal cue information associated with the vocal sample and the counterfactual synthetic vocal sample, and attribution explanation information associated with relative importance of the vocal cue information in prediction of the emotion. The processing device can also be configured to determine numeric cue differences between the vocal cue information associated with the vocal sample and the vocal cue information associated with the counterfactual synthetic vocal sample, generate cue difference relations information based on the attribution explanation information, the numeric cue differences, the vector representation of the initial prediction and the vector representation of the emotion prediction associated with the counterfactual synthetic vocal sample using a first neural network, generate a final prediction of the emotion based on the numeric cue differences, the vector representation of the initial prediction and the vector representation of the emotion prediction associated with the counterfactual synthetic vocal sample using a second neural network, and generate the explainable prediction of the emotion associated with the vocal sample based on at least the counterfactual synthetic vocal sample, the final prediction of the emotion and the cue difference relations information.

[0009] The processing device can be configured to generate the counterfactual synthetic vocal sample based on the vocal sample and the alternate emotion using a generative adversarial network.

[0010] The processing device can be configured to generate a contrastive saliency explanation based on the vocal sample, the initial prediction, and the alternate emotion using a visual explanation algorithm, and determine the vocal cue information associated with the vocal sample based on the vocal sample and the contrastive saliency explanation, and the vocal cue information associated with counterfactual synthetic vocal sample based on the counterfactual synthetic vocal sample and the contrastive saliency explanation.

[0011] The vocal cue information can be associated with one or more of a group consisting of: shrillness, loudness, average pitch, pitch range, speaking rate and proportion of pauses.

[0011a] The cue difference relations information ($\hat{\boldsymbol{r}}_w^{yy}$) can include ordinal cue difference relations information ($\hat{\boldsymbol{r}}_w^{yy}$).

[0011b] The counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{y}$), the cue difference relations information ($\hat{\boldsymbol{r}}_w^{yy}$), and the final prediction ($\hat{y}$) of the emotion are generated in a multi-task, multi-modal architecture.

[0012] Another aspect of the present disclosure provides a method for training a neural network. The method can include receiving, by a processing device, a training vector representation of the emotion associated with the vocal sample, a training vector representation of an emotion prediction associated with a counterfactual synthetic vocal sample, the counterfactual synthetic vocal sample associated with the vocal sample and an alternate emotion different from the emotion, training numeric cue difference information associated with the vocal sample and the counterfactual synthetic vocal sample and a reference emotion associated with the vocal sample. The method can also include generating,

Claims

1. A method for generating an explainable prediction of an emotion associated with a vocal sample (x), the method comprising:

receiving, by a processing device:

a vector representation ($\hat{\boldsymbol{z}}_0^{y}$) of an initial prediction ($\hat{y}_0$) of the emotion associated with the vocal sample (x);

a counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) associated with the vocal sample (x) and an alternate emotion (γ) different from the initial prediction ($\hat{y}_0$) of the emotion;

a vector representation ($\hat{\boldsymbol{z}}_0^{\gamma}$) of an emotion prediction ($\hat{\gamma}_0$) associated with the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$);

vocal cue information ($\hat{\boldsymbol{c}}^{y}, \hat{\boldsymbol{c}}^{\gamma}$) associated with the vocal sample (x) and the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$); and

attribution explanation information ($\hat{\boldsymbol{w}}_c^{y\gamma}$) associated with relative importance of the vocal cue information ($\hat{\boldsymbol{c}}^{y}, \hat{\boldsymbol{c}}^{\gamma}$) in prediction of the emotion;

determining, using the processing device, numeric cue differences ($\hat{\boldsymbol{c}}^{y\gamma}$) between the vocal cue information ($\hat{\boldsymbol{c}}^{y}$) associated with the vocal sample (x) and the vocal cue information ($\hat{\boldsymbol{c}}^{\gamma}$) associated with the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$);

generating, using the processing device, cue difference relations information ($\hat{\boldsymbol{r}}_w^{y\gamma}$) based on the attribution explanation information ($\hat{\boldsymbol{w}}_c^{y\gamma}$), the numeric cue differences ($\hat{\boldsymbol{c}}^{y\gamma}$), the vector representation ($\hat{\boldsymbol{z}}_0^{y}$) of the initial prediction ($\hat{y}_0$) and the vector representation ($\hat{\boldsymbol{z}}_0^{\gamma}$) of the emotion prediction ($\hat{\gamma}_0$) associated with the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) using a first neural network (M_r);

generating, using the processing device, a final prediction ($\hat{y}$) of the emotion based on the numeric cue differences ($\hat{\boldsymbol{c}}^{y\gamma}$), the vector representation ($\hat{\boldsymbol{z}}_0^{y}$) of the initial prediction ($\hat{y}_0$) and the vector representation ($\hat{\boldsymbol{z}}_0^{\gamma}$) of the emotion prediction ($\hat{\gamma}_0$) associated with the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) using a second neural network (M_y); and

generating, using the processing device, the explainable prediction of the emotion associated with the vocal sample (x) based on at least the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$), the final prediction ($\hat{y}$) of the emotion and the cue difference relations information ($\hat{\boldsymbol{r}}_w^{y\gamma}$).

2. The method as claimed in claim 1, wherein the step of receiving the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) comprises generating, using the processing device, the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) based on the vocal sample (x) and the alternate emotion (γ) using a generative adversarial network ($G_{\star GAN}$).

3. The method as claimed in claim 1 or 2, wherein the step of receiving the vocal cue information ($\hat{\boldsymbol{c}}^{y}, \hat{\boldsymbol{c}}^{\gamma}$) associated with the vocal sample (x) and the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) comprises:

generating, using the processing device, a contrastive saliency explanation ($\hat{\boldsymbol{\varsigma}}^{y\gamma}$) based on the vocal sample (x), the initial prediction ($\hat{y}_0$), and the alternate emotion (γ) using a visual explanation algorithm; and

determining, using the processing device, the vocal cue information ($\hat{\boldsymbol{c}}^{y}$) associated with the vocal sample (x) based on the vocal sample (x) and the contrastive saliency explanation ($\hat{\boldsymbol{\varsigma}}^{y\gamma}$), and the vocal cue information ($\hat{\boldsymbol{c}}^{\gamma}$) associated with counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) based on the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) and the contrastive saliency explanation ($\hat{\boldsymbol{\varsigma}}^{y\gamma}$).

4. The method as claimed in any one of claims 1 to 3, wherein the vocal cue information ($\hat{\boldsymbol{c}}^{y}, \hat{\boldsymbol{c}}^{\gamma}$) is associated with one or more of a group consisting of: shrillness, loudness, average pitch, pitch range, speaking rate and proportion of pauses.

5. A system for generating an explainable prediction of an emotion associated with a vocal sample (x), the system comprising a processing device configured to:

receive:

a vector representation ($\hat{\boldsymbol{z}}_0^{y}$) of an initial prediction ($\hat{y}_0$) of the emotion associated with the vocal sample (x);

a counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) associated with the vocal sample (x) and an alternate emotion (γ) different from the initial prediction ($\hat{y}_0$) of the emotion;

a vector representation ($\hat{\boldsymbol{z}}_0^{\gamma}$) of an emotion prediction ($\hat{\gamma}_0$) associated with the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$);

vocal cue information ($\hat{\boldsymbol{c}}^{y}, \hat{\boldsymbol{c}}^{\gamma}$) associated with the vocal sample (x) and the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$); and

attribution explanation information $(\widehat{\boldsymbol{w}}_c^{y\gamma})$ associated with relative importance of the vocal cue information $(\hat{\boldsymbol{c}}^{y}, \hat{\boldsymbol{c}}^{\gamma})$ in prediction of the emotion;

determine numeric cue differences $(\hat{\boldsymbol{c}}^{y\gamma})$ between the vocal cue information $(\hat{\boldsymbol{c}}^{y})$ associated with the vocal sample (x) and the vocal cue information $(\hat{\boldsymbol{c}}^{\gamma})$ associated with the counterfactual synthetic vocal sample $(\tilde{\boldsymbol{x}}^{\gamma})$;

generate cue difference relations information $(\hat{\boldsymbol{r}}_w^{y\gamma})$ based on the attribution explanation information $(\widehat{\boldsymbol{w}}_c^{y\gamma})$, the numeric cue differences $(\hat{\boldsymbol{c}}^{y\gamma})$, the vector representation $(\hat{\boldsymbol{z}}_0^{y})$ of the initial prediction $(\hat{y}_0)$ and the vector representation $(\hat{\boldsymbol{z}}_0^{\gamma})$ of the emotion prediction $(\hat{\gamma}_0)$ associated with the counterfactual synthetic vocal sample $(\tilde{\boldsymbol{x}}^{\gamma})$ using a first neural network (M_r);

generate a final prediction $(\hat{y})$ of the emotion based on the numeric cue differences $(\hat{\boldsymbol{c}}^{y\gamma})$, the vector representation $(\hat{\boldsymbol{z}}_0^{y})$ of the initial prediction $(\hat{y}_0)$ and the vector representation $(\hat{\boldsymbol{z}}_0^{\gamma})$ of the emotion prediction $(\hat{\gamma}_0)$ associated with the counterfactual synthetic vocal sample $(\tilde{\boldsymbol{x}}^{\gamma})$ using a second neural network (M_y); and

generate the explainable prediction of the emotion associated with the vocal sample (x) based on at least the counterfactual synthetic vocal sample $(\tilde{\boldsymbol{x}}^{\gamma})$, the final prediction $(\hat{y})$ of the emotion and the cue difference relations information $(\hat{\boldsymbol{r}}_w^{y\gamma})$.

6. The system as claimed in claim 5, wherein the processing device is configured to generate the counterfactual synthetic vocal sample $(\tilde{\boldsymbol{x}}^{\gamma})$ based on the vocal sample (x) and the alternate emotion (γ) using a generative adversarial network $(G_{\star GAN})$.

7. The system as claimed in claim 5 or 6, wherein the processing device is configured to:

generate a contrastive saliency explanation $(\hat{\boldsymbol{\varsigma}}^{y\gamma})$ based on the vocal sample (x), the initial prediction $(\hat{y}_0)$, and the alternate emotion (γ) using a visual explanation algorithm; and

determine the vocal cue information $(\hat{\boldsymbol{c}}^{y})$ associated with the vocal sample (x) based on the vocal sample (x) and the contrastive saliency explanation $(\hat{\boldsymbol{\varsigma}}^{y\gamma})$, and the vocal cue information $(\hat{\boldsymbol{c}}^{\gamma})$ associated with counterfactual synthetic vocal sample $(\tilde{\boldsymbol{x}}^{\gamma})$ based on the counterfactual synthetic vocal sample $(\tilde{\boldsymbol{x}}^{\gamma})$ and the contrastive saliency explanation $(\hat{\boldsymbol{\varsigma}}^{y\gamma})$.

8. The system as claimed in any one of claims 5 to 7, wherein the vocal cue information $(\hat{\boldsymbol{c}}^{y}, \hat{\boldsymbol{c}}^{\gamma})$ is associated with one or more of a group consisting of:

shrillness, loudness, average pitch, pitch range, speaking rate and proportion of pauses.

9. A method for training a neural network (M_y), the method comprising:

receiving, by a processing device:

a training vector representation of the emotion associated with the vocal sample (x);

a training vector representation of an emotion prediction associated with a counterfactual synthetic vocal sample $(\tilde{\boldsymbol{x}}^{\gamma})$, the counterfactual synthetic vocal sample $(\tilde{\boldsymbol{x}}^{\gamma})$ associated with the vocal sample (x) and an alternate emotion (γ) different from the emotion;

training numeric cue difference information associated with the vocal sample (x) and the counterfactual synthetic vocal sample $(\tilde{\boldsymbol{x}}^{\gamma})$; and

a reference emotion (y) associated with the vocal sample (x);

generating, using the processing device, an emotion prediction associated with the vocal sample (x) based on the training numeric cue difference information, the training vector representation of the emotion associated with the vocal sample (x) and the training vector representation of the emotion prediction associated with the counterfactual synthetic vocal sample $(\tilde{\boldsymbol{x}}^{\gamma})$ using the neural network (M_y);

calculating, using the processing device, a classification loss value based on differences between the emotion prediction and the reference emotion (y); and

updating, using the processing device, the neural network (M_y) to minimise the classification loss value.

10. The method as claimed in claim 9, further comprising calculating, using the processing device, attribution explanation information $(\hat{\boldsymbol{w}}_c^{y\gamma})$ with layer-wise relevance propagation of the neural network (M_y), the attribution explanation information $(\hat{\boldsymbol{w}}_c^{y\gamma})$ associated with relative importance of the vocal cue information in prediction of the emotion.

11. A method for training a neural network (M_r), the method comprising:

receiving, by a processing device:

a training vector representation of the emotion associated with the vocal sample (x);

a training vector representation of an emotion prediction associated with a counterfactual synthetic vocal sample $(\tilde{\boldsymbol{x}}^{\gamma})$, the counterfactual synthetic

vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) associated with the vocal sample (x) and an alternate emotion (γ) different from the emotion;

training numeric cue difference information associated with the vocal sample (x) and the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$);

training attribution explanation information associated with relative importance of the vocal cue information in prediction of the emotion; and

reference cue difference relations information ($\boldsymbol{r}_w^{y\gamma}$) associated with the vocal sample (x) and the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$);

generating, using the processing device, cue difference relations information based on the training attribution information, the training numeric cue differences, the training vector representation of the initial prediction and the training vector representation of the emotion prediction associated with the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) using the neural network (M_r);

calculating, using the processing device, a classification loss value based on differences between the cue difference relations information and the reference cue difference relations information ($\boldsymbol{r}_w^{y\gamma}$); and

updating, using the processing device, the neural network (M_r) to minimise the classification loss value.

12. The method as claimed in any one of claims 9 to 11, wherein the vocal cue information is associated with one or more of a group consisting of: shrillness, loudness, average pitch, pitch range, speaking rate and proportion of pauses.

13. A system for training a neural network (M_y), the system comprising a processing device configured to:

receive:

a training vector representation of the emotion associated with the vocal sample (x);

a training vector representation of an emotion prediction associated with a counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$), the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) associated with the vocal sample (x) and an alternate emotion (γ) different from the emotion;

training numeric cue difference information associated with the vocal sample (x) and the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$); and

a reference emotion associated with the vocal sample (x);

generate an emotion prediction associated with the vocal sample (x) based on the training numeric cue difference information, the training vector representation of the emotion associated with the vocal sample (x) and the training vector representation of the emotion prediction associated with the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) using the neural network;

calculate a classification loss value based on differences between the emotion prediction and the reference emotion; and

update the neural network (M_y) to minimise the classification loss value.

14. The system as claimed in claim 13, wherein the processing device is configured to calculate attribution explanation information $\left(\hat{\boldsymbol{w}}_c^{y\gamma}\right)$ with layer-wise relevance propagation of the neural network, the attribution explanation information $\left(\hat{\boldsymbol{w}}_c^{y\gamma}\right)$ associated with relative importance of the vocal cue information in prediction of the emotion

15. A system for training a neural network (M_r), the system comprising a processing device configured to:

receive:

a training vector representation of the emotion associated with the vocal sample (x);

a training vector representation of an emotion prediction associated with a counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$), the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) associated with the vocal sample (x) and an alternate emotion (γ) different from the emotion;

training numeric cue difference information associated with the vocal sample (x) and the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$);

training attribution explanation information associated with relative importance of the vocal cue information ($\hat{\boldsymbol{c}}^{y}, \hat{\boldsymbol{c}}^{\gamma}$) in prediction of the emotion; and

reference cue difference relations information ($\boldsymbol{r}_w^{y\gamma}$) associated with the vocal sample (x) and the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$);

generate cue difference relations information based on the training attribution information, the training numeric cue differences, the training vector representation of the initial prediction and the training vector representation of the emotion prediction associated with the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$) using the neural network (M_r);

calculate a classification loss value based on differences between the cue difference relations information and the reference cue difference relations information ($\boldsymbol{r}_w^{\gamma\gamma}$); and

update the neural network (M_r) to minimise the classification loss value.

16. The system as claimed in any one of claims 13 to 15, wherein the vocal cue information is associated with one or more of a group consisting of: shrillness, loudness, average pitch, pitch range, speaking rate and proportion of pauses.

17. The method as claimed in any one of claims 1 to 4, wherein the cue difference relations information ($\hat{\boldsymbol{r}}_w^{\gamma\gamma}$) includes ordinal cue difference relations information ($\hat{\boldsymbol{r}}_w^{\gamma\gamma}$).

18. The method as claimed in any one of claims 1 to 4 and 17, wherein the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$), the cue difference relations information ($\hat{\boldsymbol{r}}_w^{\gamma\gamma}$), and the final prediction ($\hat{y}$) of the emotion are generated in a multi-task, multi-modal architecture.

19. The system as claimed in any one of claims 5 to 8, wherein the cue difference relations information ($\hat{\boldsymbol{r}}_w^{\gamma\gamma}$) includes ordinal cue difference relations information ($\hat{\boldsymbol{r}}_w^{\gamma\gamma}$).

20. The system as claimed in any one of claims 5 to 8 and 19, wherein the counterfactual synthetic vocal sample ($\tilde{\boldsymbol{x}}^{\gamma}$), the cue difference relations information ($\hat{\boldsymbol{r}}_w^{\gamma\gamma}$), and the final prediction ($\hat{y}$) of the emotion are generated in a multi-task, multi-modal architecture.