Claims

1. A method of using a trained deep neural network with a triplet network architecture, the deep neural network having been trained according to a process in which dual sets of speech samples are fed through the deep neural network in combination with a cohort set of speech samples not attributed to the same speaker as the dual sets, the method comprising:

storing speech samples including a set of speaker models; and

feeding a recognition speech sample through the trained deep neural network, and verifying or identifying a user based on an output of the trained deep neural network responsive to the recognition speech sample and at least one of the speaker models.

2. The method of claim 1, further comprising

preprocessing each speech sample by:

partitioning an underlying speech signal into a plurality of overlapping windows; and

extracting a plurality of features from each of the overlapping windows.

3. The method of claim 1, further comprising:

performing a speaker verification task in which the user inputs a self-identification, and the recognition speech sample is used to confirm that an identity of the user is the same as the self-identification.

4. The method of claim 1, further comprising:

performing a speaker identification task in which the recognition speech sample is used to identify the user from a plurality of stored potential identities in association with respective speech samples.

5. The method of claim 1, further comprising: receiving a speech sample from the user as the recognition speech sample.

6. A computer-implemented method comprising:

receiving, by a computer, a recognition speech sample;

executing, by the computer, a neural network on the recognition speech sample to generate a voiceprint of the recognition speech sample, the neural network being a portion of a triplet neural architecture trained using dual sets of positive speech samples and a cohort set of negative speech samples;

comparing, by the computer, the generated voiceprint with at least one stored voiceprint; and

performing, by the computer, a speaker recognition on the recognition speech sample based upon the comparison.

7. The computer-implemented method of claim 6, wherein performing the speaker recognition comprises:

comparing, by the computer, the generated voiceprint with the at least one stored voiceprint, wherein the at least one stored voiceprint is associated with a speaker to be verified.

8. The computer-implemented method of claim 6, wherein the performing the speaker recognition comprises:

comparing, by the computer, the generated voiceprint with the at least one stored voiceprint, wherein the at least one stored voiceprint is associated with a closed set of known speakers.

9. The computer-implemented method of claim 8, wherein the closed set of known speakers is a blacklist associated with telephone fraud.

10. The computer-implemented method of claim 6, further comprising:

pre-processing, by the computer, the recognition speech sample prior to executing the neural network.

11. The computer-implemented method of claim 10, wherein the preprocessing of the recognition speech sample comprises:

segmenting, by the computer, the recognition speech sample into windows of a predetermined duration with a predetermined window shift; and

extracting, by the computer, features to be fed into the neural network from each window.

12. The computer-implemented method of claim 6, wherein performing the speaker recognition comprises at least one of:

identifying, by the computer, a speaker associated with the recognition speech sample; and

verifying, by the computer, the speaker associated with the recognition speech sample.

13. A system comprising:

a non-transitory storage medium storing a plurality of computer program instructions; and

a processor electrically coupled to the non-transitory storage medium and configured to execute the plurality of computer program instructions to perform any one of the computer-implemented methods 6-12.

14. A computer-implemented method comprising:

feeding, by a computer, a first set of one or more positive speech samples attributed to a speaker to a first feed-forward neural network to generate a first embedding vector;

feeding, by the computer, a second set of one or more positive speech samples attributed to the speaker to a second feed-forward neural network to generate a second embedding vector;

feeding, by the computer, a cohort set of negative speech samples not attributed to the speaker to a third feed-forward neural network to generate a set of embedding vectors;

calculating, by the computer, a loss function based upon the first embedding vector, the second embedding vector, and the set of embedding vectors; and

back-propagating, by the computer, the loss function to modify one or more connection weights in each of the first, second, and third feed-forward neural networks.

15. The computer-implemented method of claim 14, wherein the loss function is based upon a positive distance corresponding to a degree of similarity between the first and second embedding vectors and a negative distance corresponding to a degree of similarity between the first embedding vector and an embedding vector most similar to the first embedding vector in the set of embedding vectors.

16. The computer-implemented method of claim 15, wherein each of the degree of similarity between the first and second embedding vectors and the degree of similarity between the first embedding vector and the embedding vector most similar to the first embedding vector in the set of embedding vectors is based upon a cosine measure.

17. The computer-implemented method of claim 14, wherein the loss function is based upon an equal error rate metric associated with a degree of similarity between the first and second embedding vectors and a degree of similarity between the first embedding vector and an embedding vector most similar to the first embedding vector in the set of embedding vectors.

18. The computer-implemented method of claim 14, further comprising:

initializing, by the computer, at least one of the first, second, and third feed-forward neural network with connection weights of a deep neural network trained to perform speaker identification on a closed set of speakers.

19. A system comprising:

a non-transitory storage medium storing a plurality of computer program instructions; and

a processor electrically coupled to the non-transitory storage medium and configured to execute the plurality of computer program instructions to perform any one of the computer-implemented methods 14-17, or 18.

20. A speaker recognition device comprising:

a memory device and a processor-based device configured to perform any one of the methods 1-5.