

ABSTRACT

A speaker acquires audio data and performing processing. Audio features of the audio data are extracted. The audio features are inputted into a preset neural network model for training to obtain vector features. A first similarity value between a first vector feature and a second vector feature of a first speaker and a second similarity value between the first vector feature and a third vector feature of a second speaker are calculated. A loss function value is calculated and when is less than or equal to a preset loss function threshold, a training process of the speaker separation model is ended and parameters are updated. A two-speaker separation method, a terminal, and a storage medium are disclosed. Feature extraction capabilities of the model are enhanced, and accuracy of separation between speakers is improved, especially in long meetings and conversations.