

SPEECH PROCESSING METHOD, MODEL TRAINING METHOD AND APPARATUSES

TECHNICAL FIELD

[0001] The present disclosure relates to the field of audio processing technologies, and in particular,
5 to speech processing methods, model training methods, apparatuses, devices, and a storage medium.

BACKGROUND

[0002] Humans have an incredible ability to concentrate on a voice of their interlocutors even amid background noise, a challenge that computers struggle with. This is also often known as the cocktail party problem. A well-studied approach to this problem is to train speech separation models (SSM) on a
10 mixture of speakers, which rely on permutation invariant training (PIT) to bypass the problem of identifying or ordering its output. Target speaker extraction (TSE) is a potentially more practical approach to the problem by assuming knowledge of the target speaker through a segment of clean reference speech, which is used to extract from a mixture the speech of only the target speaker.

[0003] This background information is provided as it may be relevant to the present disclosure. No
15 admission is necessarily intended, nor should be construed, that any of the preceding information constitutes prior art against the present disclosure.

SUMMARY

[0004] In a first aspect, an embodiment of the present disclosure provides a speech processing method, including:
20 obtaining a to-be-separated speech signal, where the to-be-separated speech signal is a mixture of a target speech signal of a target speaker and at least one non-target speech signal of at least one non-target speaker;
processing, based on a reference embedding corresponding to the target speaker, the

to-be-separated speech signal with a pre-trained speech separation model to obtain the target speech signal of the target speaker, where the pre-trained speech separation model includes a separator containing a plurality of processing layers; where the plurality of processing layers comprise a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the reference embedding corresponding to the target speaker, a fused global embedding from the previous processing layer and an intermediate separation embedding, wherein the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

[0005] In a second aspect, an embodiment of the present disclosure provides a model training method, including:

acquiring a training sample set, where the training sample set includes a plurality of training samples and a sample reference embedding corresponding to a training speaker among a plurality of speakers, and each of the plurality of training samples is a mixture of speech signals from the plurality of speakers;

for a training sample in the training sample set, inputting the training sample and the sample reference embedding into a speech separation model to obtain a separation result corresponding to the training sample, and updating at least one parameter of the speech separation model based on the separation result;

where the speech separation model includes a separator containing a plurality of processing layers, and the plurality of processing layers include a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is time independent and refined with the sample reference embedding corresponding to the training speaker, a fused global embedding from the previous processing layer and an intermediate separation embedding, where the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

[0006] In a third aspect, an embodiment of the present disclosure provides a model training method,

including:

acquiring a clean sample set including a plurality of clean samples, where each of the plurality of clean samples includes clean speech data of a training speaker;

for a clean sample in the clean sample set, inputting the clean sample into a speaker
5 embedding extractor to obtain an extraction result, and updating at least one parameter of the speaker embedding extractor based on a loss value between the extraction result and a label of the clean speech data, where a speaker classification loss is excluded from the loss value.

[0007] In a fourth aspect, an embodiment of the present disclosure provides a speech processing method, including:

10 obtaining a to-be-separated speech signal collected in a meeting system, where the to-be-separated speech signal includes target speech data of a target meeting participant and non-target speech data of at least one non-target meeting participant;

processing, based on a reference embedding corresponding to the target meeting participant, the to-be-separated speech signal with a pre-trained speech separation model to obtain the target speech
15 data of the target meeting participant, where the pre-trained speech separation model includes a separator containing a plurality of processing layers; where the plurality of processing layers comprise a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer,
20 and the fused global embedding is refined with the reference embedding corresponding to the target meeting participant, a fused global embedding from the previous processing layer and an intermediate separation embedding, where the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

[0008] In a fifth aspect, an embodiment of the present disclosure provides a speech processing
25 apparatus, including:

an obtaining module, configured to obtain a to-be-separated speech signal, where the to-be-separated speech signal is a mixture of a target speech signal of a target speaker and at least one non-target speech signal of at least one non-target speaker; and

a processing module, configured to process, based on a reference embedding corresponding
30 to the target speaker, the to-be-separated speech signal with a pre-trained speech separation model to

obtain the target speech signal of the target speaker, where the pre-trained speech separation model includes a separator containing a plurality of processing layers, and the plurality of processing layers include a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the reference embedding corresponding to the target speaker, a fused global embedding from the previous processing layer and an intermediate separation embedding, where the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

10 [0009] In a sixth aspect, an embodiment of the present disclosure provides a model processing apparatus, including:

a processing module, configured to:

acquire a training sample set, where the training sample set includes a plurality of training samples and a sample reference embedding corresponding to a training speaker among a plurality of speakers, and each of the plurality of training samples is a mixture of speech signals from the plurality of speakers;

for a training sample in the training sample set, input the training sample and the sample reference embedding to obtain a separation result corresponding to the training sample;

an updating module, configured to update at least one parameter of the speech separation model based on the separation result;

20 where the speech separation model includes a separator containing a plurality of processing layers, and the plurality of processing layers include a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is time independent and refined with the sample reference embedding corresponding to the training speaker, a fused global embedding from the previous processing layer and an intermediate separation embedding, where the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

30 [0010] In a seventh aspect, an embodiment of the present disclosure provides a model training

apparatus, including:

a processing module, configured to:

acquire a clean sample set including a plurality of clean samples, where each of the plurality of clean samples includes clean speech data of a training speaker;

5 for a clean sample in the clean sample set, input the clean sample into a speaker embedding extractor to obtain an extraction result;

an updating module, configured to update at least one parameter of the speaker embedding extractor based on a loss value between the extraction result and a label of the clean speech data, where a speaker classification loss is excluded from the loss value.

10 **[0011]** In an eighth aspect, an embodiment of the present disclosure provides an electronic device, including at least one processor coupled with a memory storing a set of instructions; where the at least one processor is configured to read the set of instructions in the memory and execute the method according to the first aspect or any possible implementation of the first aspect, or the method according to the second aspect or any possible implementation of the second aspect, or the method according to
15 the third aspect or any possible implementation of the third aspect.

[0012] In a ninth aspect, an embodiment of the present disclosure provides a computer-readable medium storing computer execution instructions which, when executed by a processor, cause the processor to execute the method according to the first aspect or any possible implementation of the first aspect, or the method according to the second aspect or any possible implementation of the second
20 aspect, or the method according to the third aspect or any possible implementation of the third aspect.

[0013] In a tenth aspect, an embodiment of the present disclosure provides a computer program product including computer execution instructions which, when executed by a processor, cause the processor to execute the method according to the first aspect or any possible implementation of the first aspect, or the method according to the second aspect or any possible implementation of the second
25 aspect, or the method according to the third aspect or any possible implementation of the third aspect.

[0014] In an eleventh aspect, an embodiment of the present disclosure provides a computer program which, when executed by a processor, causes the processor to execute the method according to the first aspect or any possible implementation of the first aspect, or the second aspect or any possible implementation of the second aspect, or the method according to the third aspect or any possible
30 implementation of the third aspect.

BRIEF DESCRIPTION OF DRAWINGS

[0015] Reference will now be made, by way of example, to the accompanying drawings which show example embodiments of the present disclosure.

[0016] FIG. 1 shows a schematic illustration of a speaker fusion method.

5 [0017] FIG. 2 shows an exemplary scenario to which a speech processing method according to one or more example embodiments of the present disclosure is applied.

[0018] FIG. 3 shows a schematic flowchart of a speech processing method according to one or more example embodiments of the present disclosure.

10 [0019] FIG. 4 shows an exemplary structure of a pre-trained speech separation model according to one or more example embodiments of the present disclosure.

[0020] FIG. 5 shows a schematic flowchart of a model training method according to one or more example embodiments of the present disclosure.

[0021] FIG. 6 shows a schematic flowchart of another model training method according to one or more example embodiments of the present disclosure.

15 [0022] FIG. 7 shows a schematic flowchart of a speech processing method in meeting system according to one or more example embodiments of the present disclosure.

[0023] FIG. 8 is a schematic illustration of a speech processing method according to one or more example embodiments of the present disclosure.

20 [0024] FIG. 9 is a schematic illustration of a speech processing method according to one or more example embodiments of the present disclosure.

[0025] FIG. 10 is a schematic diagram of a histogram of GAAF with and without Cross-Entropy loss.

[0026] FIG. 11 shows a schematic structural diagram of a speech processing apparatus according to one or more embodiments of the present disclosure.

25 [0027] FIG. 12 shows a schematic structural diagram of a model training apparatus according to one or more embodiments of the present disclosure.

[0028] FIG. 13 shows a schematic structural diagram of a model training apparatus according to one or more embodiments of the present disclosure.

[0029] FIG. 14 is a structural diagram of an electronic device according to one or more embodiments

of the present disclosure.

[0030] FIG. 15 is a structural diagram of another electronic device according to one or more embodiments of the present disclosure.

DESCRIPTION OF EMBODIMENTS

5 [0031] In the following description, reference is made to the accompanying figures, which form part of the present disclosure, and which show, by way of illustration, specific aspects of embodiments of the present disclosure or specific aspects in which embodiments of the present disclosure may be used. It is understood that embodiments of the present disclosure may be used in other aspects and include structural or logical changes not depicted in the figures. The following detailed description, therefore, is
10 not to be taken in a limiting sense, and the scope of the present disclosure is defined by the appended claims.

[0032] In related arts, when training a TSE model, auxiliary information is available via an auxiliary speaker embedding extractor (or referred to as speaker embedding extractor for short), but the task is more difficult compared to SSMs because the model is only allowed a single output. Thus, the TSE
15 model has to deal with two types of errors, speaker confusion errors, where the wrong speaker is in the output, as well as separation errors, where the target speaker is not properly extracted from the non-target signal. SSMs only have to contend with the latter since PIT absolves the speaker confusion problem. Thus, SSMs are highly specialized at performing separation.

[0033] In TSE, fusion methods aim to modify intermediate representations in the SSM to guide it
20 towards the correct target output. Due to the simultaneous requirements of TSE, the fusion of speaker embeddings into SSMs would be beneficial for good performance. Various fusion methods have been studied including multiplication, concatenation and various attention-based attempts. FIG. 1 shows a schematic illustration of a speaker fusion method. Common among these existing fusion methods is that the auxiliary speaker embedding which is generated by the speaker embedding extractor and
25 provided to the model is static, i.e. regardless of the input mixture, the auxiliary speaker embedding to each Sep block is the same, as shown in FIG. 1. The Sep Block refers to a generic separation block used in an SSM. Even in the recently popular cross-attention fusion where attention is computed between speaker embeddings and the feature sequence — an inspired approach that has delivered

strong results — the speaker embedding serving as query nevertheless remains static and is not adapted to the feature sequence of the mixture.

[0034] The inventor observes that, separation is a local task where each output frame is tailored to the corresponding frame in the input. Conversely, identification is a global task because the identity of the target speaker does not change with time. The difference between global and local modeling means different model architectures are required to manage the different trade-offs, so the typical approach of relying on the SSM for both separation and identification may be sub-optimal.

[0035] In view of the above, the present disclosure provides a speech processing method, in which a to-be-separated speech signal is obtained, where the to-be-separated speech signal is a mixture of a target speech signal of a target speaker and at least one non-target speech signal of at least one non-target speaker; the to-be-separated speech signal is processed with a pre-trained speech separation model based on a reference embedding corresponding to the target speaker, to obtain the target speech signal of the target speaker, where the pre-trained speech separation model includes a separator with multiple processing layers; where for each of the multiple processing layers except for a first processing layer, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the reference embedding corresponding to the target speaker, a fused global embedding from the previous processing layer and an intermediate separation embedding obtained by the processing layer based on the output embedding from the previous processing layer. For a processing layer of the multiple processing layers of the separator, the reference embedding corresponding to the target speaker reflects the goal or objective of the identification, the intermediate separation embedding obtained by this processing layer for the output embedding from its previous processing layer reflects the identification status at this processing layer, and the fused global embedding from the previous processing layer reflects the situation of the previous processing layer (to what extent the identification has been done), therefore, the fused global embedding which is refined with the reference embedding, the fused global embedding from the previous processing layer and the intermediate separation embedding can be regarded as an adaption of the goal of the identification for this processing layer, so the reference embedding which serves as the auxiliary information for the target speaker's identification provided to the pre-trained speech separation model no longer remains static, instead, it is dynamically refined with each processing layer. This adaptability can enable the

pre-trained speech separation model to be adjusted to the nuances of the speech signal adaptably, allowing for better speaker targeting and improving the separation quality. In this way, the pre-trained speech separation model can be guided towards the target speaker gradually at each processing layer, thus the target speaker's voice can be distinguished from non-target speakers more accurately.

5 **[0036]** The speech processing method of the present disclosure method can be applied in a real-time communication system or in a meeting system especially when there are multiple participants or there is a background noise. In addition, the speech processing method of the present disclosure method can be applied in a personal audio player or smartphone, where the method can be used to enhance the clarity of a specific audio source, such as a podcast or music track, by separating it from background
10 noise. The speech processing method of the present disclosure method can be applied in an assistive listening system, a voice recognition system.

[0037] The embodiments of the present application can be used to implement speaker extraction technologies for speech, and especially can be used to implement speaker extraction for single-channel speech. FIG. 2 is a schematic diagram of an application scenario according to an embodiment of the
15 present application. As shown in FIG. 2, in a meeting, multiple participants A, B, and C can use the same speech input device, such as a microphone. The speech input device transmits acquired single-channel to-be-separated speech to a processing device, which then processes the to-be-separated speech (signal). The processing device performs speech identification on the to-be-separated speech to extract a target participant's speech from the to-be-separated speech.

20 **[0038]** FIG. 3 shows a schematic flowchart of a speech processing method according to one or more example embodiments of the present disclosure. The method may be applied in the scenario shown in FIG. 2, and may also be applied in other scenarios where a speaker's voice needs to be extracted from a mixture of several speakers' voices. Illustratively, the speech processing method of the present disclosure can be implemented by a speech processing device provided by the present disclosure, and
25 the speech processing device can be implemented by any software and/or hardware. The speech processing device can be the processing device shown in FIG. 2. For example, the speech processing device can be: a tablet computer, a mobile phone (such as a folding screen mobile phone, a large screen mobile phone, etc.), a wearable device, a vehicle-mounted device, an augmented reality (AR)/virtual reality (VR) equipment, a laptop computer, a ultra-mobile personal computer (UMPC), a netbook, a
30 personal digital assistant (PDA), a smart TV, a smart screen, a HD TV, a 4K TV, a smart speaker, a

smart projector and other Internet of Things (IOT) devices. The specific types of the speech processing device are not limited in the embodiments of the present disclosure. For example, the method provided in the embodiments of the present disclosure can be implemented through an all-in-one conference system, or via a terminal such as a smartphone, a computer, a tablet device, etc. In a possible implementation, the terminal can send the to-be-processed speech signal to a server, which then uses the method provided in the embodiments of the present disclosure to obtain a speaker separation result and feeds back the result to the terminal. As shown in FIG. 3, the method can include the following steps.

[0039] S301, obtain a to-be-separated speech signal, where the to-be-separated speech signal is a mixture of a target speech signal of a target speaker and at least one non-target speech signal of at least one non-target speaker.

[0040] In the embodiment, the to-be-separated speech signal is obtained. The to-be-separated speech signal is a mixture of a target speech signal of a target speaker and at least one non-target speech signal of at least one non-target speaker. In a possible implementation, the to-be-separated speech signal may be obtained through a speech input apparatus. For example, the to-be-separated speech signal can be obtained through a microphone, or a personal device like a smartphone or tablet, or any other audio recording setup. The mixture may be typically represented as a waveform or a series of digital samples, which carries acoustic information of at least part of the speakers. The number of non-target speakers is not limited in the embodiments of the present disclosure.

[0041] S302, process the to-be-separated speech signal with a pre-trained speech separation model based on a reference embedding corresponding to the target speaker, to obtain the target speech signal of the target speaker, where the pre-trained speech separation model includes a separator containing a plurality of processing layers, and the plurality of processing layers include a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers (i.e., the first one among the multiple processing layers), the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the reference embedding corresponding to the target speaker, a fused global embedding from the previous processing layer and an intermediate separation embedding, where the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

[0042] In the embodiment, after obtaining the to-be-separated speech signal, the processing device processes the to-be-separated speech signal with a pre-trained speech separation model based on a reference embedding corresponding to the target speaker, to obtain the target speech signal of the target speaker. The reference embedding may be obtained from a reference speech signal of the target speaker, so it may be a representation that corresponds to the target speaker. The reference embedding, acting as a guide for the model, may be used for the pre-trained speech separation model to identify and focus on the speech of the target speaker. The pre-trained speech separation model is used to process the mixed speech signal, and the pre-trained speech separation model includes a separator with multiple processing layers. Each processing layer within the separator is responsible for separation based on the separation result from its previous processing layer (which is represented as output embedding from the previous processing layer), and outputs an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer. The fused global embedding is continuously refined throughout the processing layers as it is combined with the reference embedding, a fused global embedding from the previous layer and the intermediate separation embedding generated for the output embedding from the previous layer. For each processing layer, the output embedding from its previous processing layer serves as the input embedding of this processing layer, and may provide context from the preceding stage of separation.

[0043] It should be noted that normally there are several processing layers in the separator, and the first one generally takes a speech encoding embedding obtained through encoding of the to-be-separated speech signal as its input since there is actually no previous processing layer for the first processing layer, in this regard, for the first processing layer, the output embedding from its previous processing layer would simply refer to the speech encoding embedding. The first processing layer receives the speech encoding embedding from, e.g., an encoder of the pre-trained speech separation model, and also receives the reference embedding corresponding to the target speaker from, e.g., a speaker embedding extractor, and then generates a fused global embedding and passes this fused global embedding to the next processing layer.

[0044] For any one of remaining processing layers, it receives, from the previous processing layer, the fused global embedding and the output embedding obtained at the previous processing layer, and also receives the reference embedding corresponding to the target speaker from, e.g., a speaker embedding extractor, then it refines those embeddings together to generate a fused global embedding

and passes this fused global embedding to the next processing layer. In this way, the adjustment for the identification task to be completed by each processing layer is realized through continuous refinement of the fused global embedding through layers.

[0045] As described above, the reference embedding is time independent (which means this representation is irrelevant to time), and the fused global embedding obtained at each processing layer is also time independent, and this time independent fused global embedding serves as auxiliary information for the generation of the output embedding at each processing layer, so such auxiliary information is no longer static, it is adapted to the previous separation result and thus the continuous passing of the time independent fused global embedding through layers guides the whole speech separation model to a more accurate identification result.

[0046] According to the method provided by the embodiments of the present disclosure, a to-be-separated speech signal is obtained, the to-be-separated speech signal is processed with a pre-trained speech separation model based on a reference embedding corresponding to the target speaker, to obtain the target speech signal of the target speaker. By using a reference embedding that corresponds to the target speaker, the target speaker's voice can be distinguished from non-target speakers more accurately. The pre-trained speech separation model includes a separator with multiple processing layers, and for each of the multiple processing layers except for a first processing layer, the processing layer is configured to output an output embedding based on a fused global embedding, and an output embedding from a previous processing layer of the processing layer. The fused global embedding is refined with the reference embedding corresponding to the target speaker, a fused global embedding from the previous processing layer and an intermediate separation embedding obtained by the processing layer based on the output embedding from the previous processing layer. For a processing layer of the multiple processing layers of the separator, the reference embedding corresponding to the target speaker reflects the goal or objective of the identification, the intermediate separation embedding obtained by this processing layer for the output embedding from its previous processing layer reflects the identification status at this processing layer, and the fused global embedding from the previous processing layer reflects the situation of the previous processing layer (to what extent the identification has been done), therefore, the fused global embedding which is refined with the reference embedding, the fused global embedding from the previous processing layer and the intermediate separation embedding can be regarded as an adaption of the goal of the identification for

this processing layer, so the reference embedding which serves as the auxiliary information for the target speaker's identification provided to the pre-trained speech separation model no longer remains static, instead, it is dynamically refined with each processing layer. This adaptability can enable the pre-trained speech separation model to be adjusted to the nuances of the speech signal adaptably, allowing for better speaker targeting and improving the separation quality. In this way, the pre-trained speech separation model can be guided towards the target speaker gradually at each processing layer, thus the target speaker's voice can be distinguished from non-target speakers more accurately.

[0047] In the following, the detailed implementations will be described with reference to an exemplary structure of the pre-trained speech separation model shown in FIG. 4. As shown in FIG. 4, in addition to the separator, the speech separation model may optionally include an encoder and a decoder.

[0048] Generally, the processing device may obtain a speech encoding embedding through the encoder based on the to-be-separated speech signal. For an i -th processing layer among the multiple processing layers of the separator, the processing device may obtain an i -th intermediate separation embedding based on a first input of the i -th processing layer, may generate an i -th fused global embedding based on a second input of the i -th processing layer and an i -th global embedding of the i -th intermediate separation embedding, and may output an i -th output embedding based on the i -th intermediate separation embedding and the i -th fused global embedding; where i is a positive integer not greater than N , N is a number of the multiple processing layers. When i equals to 1, the first input of the i -th processing layer is the speech encoding embedding, and the second input of the i -th processing layer is the reference embedding; when i is greater than 1 but not greater than N , the first input of the i -th processing layer is an $(i-1)$ -th output embedding from an $(i-1)$ -th processing layer, and the second input of the i -th processing layer is the reference embedding and an $(i-1)$ -th fused global embedding from the $(i-1)$ -th processing layer. The processing device may obtain the target speech signal of the target speaker through the decoder based on an output embedding output by an N -th processing layer among the multiple processing layers and the speech encoding embedding.

[0049] The speech separation model may incorporate an encoder and a decoder to process and extract the target speech signal from the to-be-separated speech signal. The encoder and the decoder are conceptual components which can be implemented in software or hardware, which is not limited here.

The encoder and decoder may be components of the speech separation model that are respectively

responsible for transforming a raw speech signal into a suitable representation for processing and then reconstructing the separated speech signal. The encoder may be a neural network layer or a sequence of layers designed to convert a raw speech signal (or the to-be-separated speech signal) into a compact and informative embedding. It may operate on the input speech waveform, which may be in the form of a time-domain signal or a spectrogram (which may be a representation in speech processing that shows the spectrum of frequencies as a function of time). The encoder can utilize various techniques such as convolutional layers, recurrent layers (like LSTM or GRU), or attention mechanisms to capture relevant features from the speech signal. The output of the encoder may be a speech encoding embedding that summarizes the input speech in a way that can be effectively used by a subsequent processing layer.

[0050] The decoder may be another neural network component that takes the output from the N-th processing layer (i.e., the last processing layer) of the speech separation model and transforms it back into a speech signal that approximates the original speech of the target speaker. The decoder may perform an inverse process of the encoder operation, taking the embedded representation and reconstructing a time-domain waveform. The decoder may use techniques such as transposed convolutions, recurrent layers, or a vocoder (which is a neural network specifically designed for generating high-fidelity speech from spectrograms or embeddings) to convert the embeddings back into a speech signal. The output of the decoder may be the target speech signal of the target speaker, which may be a time-domain waveform that can be played back or further processed.

[0051] The encoder may be used to obtain a speech encoding embedding from the to-be-separated speech signal, and the decoder may be used to obtain the target speech signal of the target speaker based on an output embedding output by an N-th processing layer among the multiple processing layers and the speech encoding embedding. The reference embedding may be obtained from a reference speech signal of the target speaker, so the reference embedding corresponds to the target speaker. The reference embedding may act as a guide for the model to identify and focus on the target speaker's voice within the mixture. The to-be-separated speech signal, including the target speaker's voice along with other non-target speakers' voices, may be processed to obtain a speech encoding embedding, which may be used to represent an initial and unseparated mixture of voices. The pre-trained speech separation model may include multiple processing layers, and each layer may perform operations to refine the separation of the target speaker's voice. For each processing layer (that is, the i-th layer), an

intermediate separation embedding may be obtained, and the intermediate separation embedding is based on the first input to the layer, which may be either the initial speech encoding embedding for the first layer or the output embedding from the previous layer for subsequent layers. The second input to the i -th processing layer may be used in conjunction with the i -th global embedding of the intermediate separation embedding to generate an i -th fused global embedding. The i -th fused global embedding may be refined based on the reference embedding, the $(i-1)$ -th fused global embedding and the intermediate separation embedding (the $(i-1)$ -th fused global embedding would be omitted for the first processing layer). The i -th processing layer may output an i -th output embedding. This output embedding may be based on both the i -th intermediate separation embedding and the i -th fused global embedding, combining local speech information with the global speaker information.

[0052] The first input to the i -th processing layer may depend on the layer's position in the pre-trained speech separation model. For the first layer (that is, when $i=1$), the first input is the speech encoding embedding, and the second input is the reference embedding. For layers greater than the first (that is, when $i>1$), the first input is the output embedding from the previous $(i-1)$ -th layer, and the second input may include both the reference embedding and the $(i-1)$ -th fused global embedding. The target speech signal of the target speaker may be obtained based on the output embedding from the last processing layer and the initial speech encoding embedding. The final output may be a refined representation that isolates the target speaker's voice from the mixture.

[0053] In a possible implementation, the output embedding output by an N -th processing layer among the multiple processing layers and the speech encoding embedding may be a mask of the target speaker. The mask may be a time-frequency representation where each entry may indicate a degree to which a corresponding time-frequency bin of the speech signal is attributed to the target speaker. The masked representation may be passed through a decoder, which converts it back into a time-domain speech signal. The decoder might use the initial speech encoding embedding to help reconstruct a high-fidelity speech signal that closely resembles the original speech of the target speaker.

[0054] In this way, each processing layer can build upon the outputs of the previous layers, which can refine the global embeddings, and gradually improve the extraction of the target speaker's voice. The method leverages the reference embedding to maintain focus on the target speaker throughout the separation process, which can be beneficial for achieving high performance in audio mixtures.

[0055] In a possible implementation, each of the multiple processing layers includes a separation

part and a fusion part. The processing device may input the first input of the i -th processing layer into a separation part of the i -th processing layer to obtain the i -th intermediate separation embedding, and the processing device may input the i -th intermediate separation embedding and the second input of the i -th processing layer into a fusion part of the i -th processing layer to obtain the i -th fused global embedding.

5 The processing device may output the i -th fused global embedding to the separation part of the i -th processing layer, and the processing device may output the i -th output embedding by the separation part of the i -th processing layer based on the i -th intermediate separation embedding and the i -th fused global embedding from the fusion part of the i -th processing layer. The processing may be separated into distinct parts (that is, the separation part and the fusion part) within of the multiple processing
10 layers. The separation and fusion parts can potentially be processed in parallel, which may improve computational efficiency and reduce the overall processing time. The separation part may be responsible for performing the initial separation of the speech signal, by taking an input and attempting to extract features that are specific to the target speaker. The fusion part may be responsible for integrating auxiliary information, such as a speaker/reference embedding, with the output of the
15 separation part to refine the separation process. The first input of the i -th processing layer may be processed by the separation part of the i -th processing layer, and the processing result may be an intermediate separation embedding, which may be a feature representation that captures the separation of the target speaker's speech at this stage of processing. The second input of the i -th processing layer along with the intermediate separation embedding, may be used by the fusion part of the i -th processing
20 layer. The fusion part may integrate the i -th intermediate separation embedding and the second input to generate an i -th fused global embedding, which may be a refined representation that combines the separation-specific features with the auxiliary speaker information. The i -th fused global embedding may be then output and provided to the separation part of the same i -th layer. The separation part of the i -th processing layer may use the i -th intermediate separation embedding and the i -th fused global
25 embedding to produce the final output for the i -th processing layer. The i -th output embedding may be a refined feature representation that may be used as the input to the next layer (that is, the $(i+1)$ -th processing layer) or, in the case of the last processing layer, may be used to obtain the target speech signal through the decoder.

[0056] Separation of the functionality into distinct “separation” and “fusion” parts within each layer
30 allows for a modular approach, the development, optimization, and potential reuse of these components

can be simplified. By obtaining intermediate embeddings and fusing them with global embeddings, each processing layer can incrementally refine the representation of the target speaker's voice.

[0057] In a possible implementation, as shown in FIG. 4, the fusion part of the i -th processing layer may include a global extracting sub-layer, an adaption sub-layer and a fusion sub-layer. The processing device may obtain the i -th global embedding of the i -th intermediate separation embedding through the global extracting sub-layer, may adapt the second input of the i -th processing layer and the i -th global embedding of the i -th intermediate separation embedding to processing at the fusion sub-layer through the adaption sub-layer, and the processing device may fuse the adapted second input of the i -th processing layer and the i -th adapted global embedding of the i -th intermediate separation embedding through the fusion sub-layer to obtain the i -th fused global embedding, where a size of the i -th fused global embedding is matched with a size of the i -th global embedding of the i -th intermediate separation embedding. It should be noted that in FIG. 4, only the structure of the second processing layer among the multiple processing layers is shown in detail, but other processing layers also have their specific fusion parts and separation parts and may function as described above, which is not illustrated in the figure for brevity.

[0058] The global extracting sub-layer may be responsible for deriving the i -th global embedding from the i -th intermediate separation embedding. The global embedding may capture the global, time-independent features of the speech signal that are relevant for speaker identification and separation.

[0059] The adaption sub-layer may take the second input of the i -th processing layer (which may include the reference embedding and, for $i > 1$, the $(i-1)$ -th fused global embedding) and the i -th global embedding from the previous step. The adaption sub-layer may act as an adapter, which may convert the different representations (the second input and the i -th global embedding) into a unified format that can be effectively combined by the fusion sub-layer. In a possible implementation, the adaption sub-layer may be part of an embedding mixer or a separate component that is responsible for preparing the inputs for the fusion sub-layer. This adaptation process may involve operations such as rearranging, scaling, normalization, enhancing, filtering, or other transformations to make the representations compatible and useful for the subsequent fusion step. For example, the adaption sub-layer may rearrange the elements of the input vectors/embeddings to align them in a way that is more conducive to the fusion process, and the rearranging operation may involve changing the order of elements or the

structure of the representation to match the expected input format of the fusion sub-layer. The adaption sub-layer may enhance certain features of the inputs that are particularly relevant for the separation task, and the enhancing operation may involve amplifying specific frequency components, emphasizing particular temporal features, or highlighting aspects of the speaker's voice that are important for identification and separation. For another example, the adaption sub-layer may also filter out noise or irrelevant information from the inputs, and the filtering operation may involve removing frequency components that are not useful for the separation, suppressing background noise, or eliminating artifacts that could interfere with the fusion process.

[0060] In a possible implementation, the adaption sub-layer is a single convolutional network. A convolutional network can share weights across different parts of the input, which can lead to more efficient learning and generalization. In addition, the convolutional network may behave well at mapping input features to a higher-dimensional space where the separation of different speakers can be more distinct.

[0061] In a possible implementation, the processing device may perform at least one of a rearrangement, filtering or enhancing operation on elements of the i -th global embedding of the i -th intermediate separation embedding and each embedding in the second input of the i -th processing layer to obtain the adapted second input of the i -th processing layer and the i -th adapted global embedding of the i -th intermediate separation embedding. The elements of the i -th global embedding may be individual components or features within a higher-level representation that captures global aspects of the input data at the i -th processing layer. The elements of the i -th global embedding may be at least one of the i -th global embedding: a mel-frequency cepstral coefficient (MFCC), a formant frequency, a voice/unvoiced indicator, loudness or energy, a harmonic-to-noise ratio (HNR), a spectral feature, a temporal feature, contextual information, etc. By adapting the embeddings through at least one of rearrangement, filtering, and enhancement, the speech separation model can better leverage the information from the intermediate separation embedding and the auxiliary input(s) to improve the speech separation performance.

[0062] The fusion sub-layer may be responsible for integrating the adapted second input and the i -th adapted global embedding. The fusion operation may combine the global speaker information with the intermediate separation embedding to produce the i -th fused global embedding. The size of the i -th fused global embedding may be matched with the size of the i -th global embedding of the i -th

intermediate separation embedding to ensure compatibility and facilitate the subsequent processing within the separation part of the i -th layer. Here the size refers to the embedding size of an embedding. For example, the dimension of input and output embeddings of the separation part is three, including time, chunk dimension and embedding size, but in the fusion part, the dimensions of the $(i-1)$ -th fused global embedding, the i -th global embedding of the i -th intermediate separation embedding, and the output i -th fused global embedding is one, simply including the dimension of the embedding size. Then the i -th fused global embedding is obtained in such a way that its embedding size is matched with the embedding size of the i -th global embedding of the i -th intermediate separation embedding.

[0063] The global extracting sub-layer may capture the global features, the adaption sub-layer may prepare these features for integration, and the fusion sub-layer may combine them to enhance the separation process. The structured approach can allow the speech separation model to progressively refine the separation of the target speaker's speech throughout the multiple processing layers.

[0064] In a possible implementation, the processing device may input the i -th intermediate separation embedding into the global extracting sub-layer and determine the last element from the i -th intermediate separation embedding as the i -th global embedding. For example, as described above, the dimension of the input embedding of the separation part is three, including time, chunk dimension and embedding size, so as one possible implementation, a global extracting sub-layer of the fusion part, chunk pooling may be performed on the intermediate separation embedding to obtain an embedding of the embedding size for each chunk, and then a last element of each chunk is selected as the global embedding. The selection could be the result of a temporal dimension reduction and a chunk dimension reduction, so the processing device may focus on the most recent information. When chunks of a speech signal are created with overlap, the last element of one chunk is also the first element of the subsequent chunk. This overlap means that the last element may inherently include information about the end of the previous chunk's context and the beginning of the next chunk's context, thus capturing a form of global information across the sequence. In transformer models, an attention mechanism can be used to give different weights to different parts of the input sequence. The last element can be adapted by the attention mechanism to store and reflect global information because it can be emphasized during the attention process, allowing it to represent the entire chunk effectively. In addition, by determining the last element from the i -th intermediate separation embedding as the i -th global embedding, an existing element is selected rather than creating a new representation or combining elements through

parameterized operations. By selecting the last element of the intermediate separation embedding, global information can be captured adaptively and efficiently, without the need for additional parameters or complex mechanisms, which can help extracting a good global embedding for the speech separation model.

5 **[0065]** In a possible implementation, the processing device may concatenate the i-th adapted global embedding of the i-th intermediate separation embedding and each embedding in the adapted second input of the i-th processing layer, and may extract a number of weighted elements in the concatenated embedding as the i-th fused global embedding, where the number of weighted elements in the concatenated embedding is as same as a size of the i-th global embedding of the i-th intermediate separation embedding. The processing device may concatenate the i-th adapted global embedding with
10 each embedding within the adapted second input, to obtain a single, combined representation that may include both the global features and the adapted input features. From the concatenated embedding, a number of weighted elements may be extracted. These weighted elements may be chosen based on their significance or relevance, which can be determined through a weighting scheme (such as, a set of
15 trainable weights, an attention mechanism, a sigmoid activation function, etc.). The number of weighted elements extracted may be equal to the size of the i-th global embedding of the i-th intermediate separation embedding, which can ensure that the fused global embedding maintains the same size as the original global embedding, allowing it to be effectively used when being feed back to the separation part of the same processing layer and also when being passed to a next processing layer.
20 The collection of weighted elements may form the i-th fused global embedding, which may be a refined representation that combines global information with the adapted features from the second input.

[0066] In a possible implementation, the processing device may align and average the i-th adapted global embedding of the i-th intermediate separation embedding and each embedding in the second
25 input of the i-th processing layer, and obtain the i-th fused global embedding based on an averaged embedding, the averaged embedding is an average of the i-th adapted global embedding of the i-th intermediate separation embedding and each embedding in the second input of the i-th processing layer. The adapted global embedding from the i-th intermediate separation embedding may be aligned with each embedding within the adapted second input, which can ensure that the embeddings are structured
30 in a way that allows for element-wise interaction or integration. The aligned embeddings are then

averaged, which is an effective way to combine information from multiple sources. The averaging operation may help to create a fused representation that balances the global speaker information with the specific details from the adapted second input.

[0067] The fusion sub-layer may be an embedding mixer, and may be used to combine the outputs of embeddings (that is, the adapted second input of the i -th processing layer and the i -th adapted global embedding of the i -th intermediate separation embedding). There may be many ways for implementing such fusion sub-layer, here two examples are given for illustration:

[0068] Example 1: when the input embeddings have different (embedding) sizes and need to be combined into a single embedding, the input embeddings may be adapted through the adaption sub-layer and the adapted embeddings can be concatenated and a number of weighted elements in the concatenated embedding can be extracted. For example, if there are one 256-size embedding representing the global representation of the i -th processing layer, another 256-size embedding from a previous layer (that is the $(i-1)$ -th processing layer), and a 512-size embedding, they can be concatenated into a single 1024-size vector. The concatenated vector can be then weighted to emphasize the most significant features. The Audio Mixer, acting as a linear layer, can apply these weights to select the most important 256 features as the output, i.e., the fused global embedding.

[0069] Example 2: when all input embeddings are of the same (embedding) sizes, a straightforward averaging can be applied without the need for concatenation. The input embeddings may be adapted through the adaption sub-layer and the adapted embeddings, and then can be aligned, so they can be combined directly. For example, an average of the aligned embeddings can be taken to create a single vector that includes the fused information.

[0070] It should be noted that “dimension” refers to the number of independent features or axes in a vector space. “Size” refers to the length of a vector within a single dimension, such as the number of elements in a one-dimensional vector. The inputs and the outputs of the Audio Mixer can have the same or different sizes, which is not limited here. Besides, the division of the fusion part into a global extracting sub-layer, an adaption sub-layer and a fusion sub-layer is illustrative rather than restrictive, other divisions may be possible, as long as functions as described above for the sub-layers are realized.

[0071] In this way, information from both the adapted global embedding and the adapted second input can be integrated, a more comprehensive representation that can capture a wider range of speaker characteristics can be generated. In addition, the size of the fused global embedding can be ensured to

match the size of the original global embedding, which maintains consistency in the dimensionality of the embeddings throughout the processing layers.

[0072] In a possible implementation, the separation part of the i -th processing layer may be a single-path global modulation (SPGM) block. The SPGM block can be designed to modulate the speech representation in a way that emphasizes the target speaker’s features, which can lead to more effective separation. In addition, the SPGM block may be a straightforward architecture which may be less prone to overfitting and easier to optimize compared to more complex, multi-path architectures.

[0073] In a possible implementation, the processing device may process a reference speech signal with a pre-trained speaker embedding extractor to obtain the reference embedding corresponding to the target speaker, where the reference speech signal includes clean speech data of the target speaker, and the pre-trained speaker embedding extractor is obtained from a training for speech extraction during which a speech classification loss is skipped. The reference speech signal may include clean and non-mixed speech data, ensuring that the reference embedding is of high quality and accurately represents the target speaker’s voice. During the training of the speaker embedding extractor, a speech classification loss may be intentionally skipped. This training strategy can make the model focus on extracting robust speaker features rather than classifying speaker identities.

[0074] In many TSE models in the related art, speaker embedding extractors produce an embedding that is jointly trained on both a speaker classification task and on the TSE task. The goal of the secondary speaker classification loss is to guide the model towards more discriminative embeddings.

The speaker classifier is a linear layer over the speaker embeddings producing speaker classes. The loss is computed using Cross Entropy (L_{CE}) and added to the main SI-SDR loss. However, L_{CE} may be unnecessary because incorrect speaker classification also penalizes the main SI-SDR loss, and having a single loss prevents potential gradient conflicts. The exclusion of the speech classification loss may simplify the training objective, which can lead to more efficient and stable optimization of the model.

The use of clean speech data for the reference signal may ensure that the obtained reference embedding is a high-fidelity representation of the target speaker’s voice. In addition, skipping the speech classification loss could help simplifying the training objective, focusing the model on the separation task rather than on classifying speakers.

[0075] In a possible implementation of the first aspect, the pre-trained speaker embedding extractor is a pre-trained Asymmetric Cross Attention (ACA) net with a frequency-domain encoder. In this way,

speaker embeddings can be efficiently extracted without requiring excessive computational resources. In addition, since the cross-attention mechanism in ACA-Net has an asymmetric nature, the capture of both global and local speaker features can be allowed, which can be beneficial for speaker discrimination. ACA-Net can be chosen for the auxiliary speaker embedding extractor because it is
5 lightweight and has good performance on a WSJ0-2mix dataset speaker verification problem, outperforming popular speaker verification models like ECAPATDNN. This model obtains time-independent global information using asymmetrical cross-attention, which requires significantly fewer parameters than typical speaker verification models. This structure sets up a strong time-independence prior which is a desirable trait for a speaker embedding extractor. The original
10 frequency-domain encoder can be used by ACA-Net while using a time-domain encoder for the SSM.

[0076] The speech processing method of the present disclosure will be described in conjunction with the structure of the pre-trained speech separation model in the following.

[0077] In a possible implementation, each of the multiple processing layers may include a separating block (similar to the separation part as described above) and a fusion block (similar to the fusion part as
15 described above). A fusion block of an i -th processing layer may be configured to output an i -th fused global embedding, to an $(i+1)$ -th processing layer and a separating block of the i -th processing layer, based on an i -th global embedding of an i -th intermediate separation embedding output by the separating block of the i -th processing layer and the reference embedding vector corresponding to the target speaker. The separating block of the i -th processing layer may be configured to output an i -th
20 output embedding based on the i -th intermediate separation embedding and the i -th fused global embedding. Here i is a positive integer greater than 0 not greater than N , N is a number of the multiple processing layers, for i being 1, the i -th intermediate separation embedding may be obtained based on a speech encoding embedding of the to-be-separated speech signal; for i being greater than 1, the outputting of the i -th fused global embedding is further based on an $(i-1)$ -th fused global embedding
25 from an $(i-1)$ -th processing layer, and the i -th intermediate separation embedding is obtained based on an $(i-1)$ -th output embedding output by a separating block of the $(i-1)$ -th processing layer.

[0078] In a possible implementation, for i being 1, the fusion block of the i -th processing layer may be configured to perform at least one of a rearrangement, filtering or enhancing operation on the reference embedding and the i -th global embedding of the i -th intermediate separation embedding
30 respectively and fuse the rearranged/filtered/enhanced reference embedding and i -th global embedding

of the i-th intermediate separation embedding to obtain the i-th fused global embedding.

[0079] In a possible implementation, the fusion may include concatenating the rearranged/filtered/enhanced reference embedding and i-th global embedding of the i-th intermediate separation embedding to obtain a concatenated embedding and extracting a number of weighted elements in the concatenated embedding as the i-th fused global embedding with a size being the same as a size of the i-th global embedding of the i-th intermediate separation embedding.

[0080] In a possible implementation, for i being greater than 1, the fusion block of the i-th processing layer may be configured to perform at least one of a rearrangement, filtering or enhancing operation on the reference embedding, the (i-1)-th fused global embedding, and the i-th global embedding of the i-th intermediate separation embedding respectively and fuse the rearranged/filtered/enhanced reference embedding, (i-1)-th fused global embedding and i-th global embedding of the i-th intermediate separation embedding to obtain the i-th fused global embedding.

[0081] In a possible implementation, the fusion may include concatenating the rearranged/filtered/enhanced reference embedding, (i-1)-th fused global embedding and i-th global embedding of the i-th intermediate separation embedding to obtain a concatenated embedding and extracting a number of weighted elements in the concatenated embedding as the i-th fused global embedding with a size being the same as a size of the i-th global embedding of the i-th intermediate separation embedding.

[0082] FIG. 5 shows a schematic flowchart of a model training method according to one or more example embodiments of the present disclosure. The model training method can be used to train the above-mentioned pre-trained speech separation model according to one or more embodiments of the present disclosure. The method may be implemented by an apparatus, such as a speech processing apparatus, a (cloud) server or other devices, such as a chip which has similar function. In a possible implementation, the apparatus may be a processing device. As shown in FIG. 5, the method can include the following steps.

[0083] S501, acquire a training sample set, where the training sample set comprises a plurality of training samples and a sample reference embedding corresponding to a training speaker among a plurality of speakers, and each of the multiple training samples is a mixture of speech signals from the plurality of speakers.

[0084] S502, for a training sample in the training sample set, input the training sample and the

sample reference embedding into a speech separation model to obtain a separation result corresponding to the training sample, and updating at least one parameter of the speech separation model based on the separation result.

[0085] The speech separation model includes a separator containing a plurality of processing layers and the plurality of processing layers include a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers except for a first processing layer, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the sample reference embedding corresponding to the training speaker, a fused global embedding from the previous processing layer and an intermediate separation embedding, where the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

[0086] In the embodiment, a training sample set including a plurality of training samples and a sample reference embedding corresponding to a training speaker among a plurality of speakers are acquired, each of the plurality of training samples is a mixture of speech signals from the plurality of speakers. For the training sample in the training sample set, the training sample and the sample reference embedding are input into the speech separation model to obtain a separation result corresponding to the training sample, and then at least one parameter of the speech separation model is updated based on the separation result, the above process continues iteratively until a trained speech separation model is obtained. The acquiring of the training sample set would be performed once for each training epoch, and in each training iteration of a training epoch, the inputting of the training sample(s) and the sample reference embedding into the speech separation model would be performed. For each training iteration, there would one or more training samples used for training, which is not limited herein, and the following description would be made by taking one training sample per training iteration as an example, but would also be applicable for the case of more than one training sample per training iteration.

[0087] Here the training speaker is a target speaker whose voice is expected to be output by the speech separation model. The training sample is a mixture of speech signals from multiple speakers (including the training speaker and other speakers). For example, the training samples used in different iterations would be different mixtures of speech signals from the multiple speakers, and the separation

result obtained in each iteration would be a speech signal of a speaker (which could be the training speaker if the speech separation model works correctly or not the training speaker if the speech separation model works incorrectly). The speech separation model includes a separator containing a plurality of processing layers. The plurality of processing layers include a first processing layer and
5 remaining processing layers. For a processing layer of the remaining processing layers: the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the separator; and the fused global embedding is time-independent and is refined with the sample reference embedding corresponding to the training speaker, a fused global embedding from the previous processing layer and an intermediate separation
10 embedding, where the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

[0088] In each training iteration, after obtaining the separation result, at least one parameter of the speech separation model is updated based on the separation result corresponding to a training sample used in this training iteration. For example, since the separation result obtained in each training
15 iteration would be a speech signal of a speaker, so in each training iteration, the obtained speech signal would be compared with ground truth data (clean speech data/clean speech signal of the training speaker in this training iteration), and the at least one parameter could be updated based on a result of the comparison.

[0089] The speech separation model can fuse three different embeddings (sample reference embedding, global embedding and fused global embedding) and outputs a refined embedding of a size
20 appropriate for a primary SSM. The sample reference embedding may be speaker-specific features corresponding to the training speaker. The global embedding is time-independent and refined at each layer of the model to better capture the characteristics of the speaker. The fused global embedding combines the sample reference embedding and the intermediate separation embedding to guide the
25 model towards the correct target output for separation.

[0090] By using the model training method, a speech separation model with an improved capability of separating the target speaker's voice from a mixture of speech signals can be obtained. In this way, the target speaker extraction can be improved by focusing on refining the separation process through multiple processing layers, mixed signals can be better handled, and the efficiency of model training
30 can be improved.

[0091] FIG. 6 shows a schematic flowchart of another model training method according to one or more example embodiments of the present disclosure. The model training method can be used to train the above-mentioned pre-trained speaker embedding extractor according to one or more embodiments of the present disclosure. The method may be implemented by an apparatus, such as a speech processing apparatus, a (cloud) server or other devices, such as a chip which has similar function. In a possible implementation, the apparatus may be a processing device. As shown in FIG. 6, the method can include the following steps.

[0092] S601, acquire a clean sample set including a plurality of clean samples, where each of the plurality of clean samples includes clean speech data of a training speaker.

[0093] S602, for a clean sample in the clean sample set, input a clean sample into a speaker embedding extractor to obtain an extraction result.

[0094] S603, update at least one parameter of the speaker embedding extractor based on a loss value between the extraction result and a label of the clean speech data, where a speaker classification loss is excluded from the loss value.

[0095] In the embodiment, a clean sample set including a plurality of clean samples is acquired, then a clean sample in the clean sample set, which includes clean speech data of a training speaker, is input into the speaker embedding extractor to obtain an extraction result, and at least one parameter of the speaker embedding extractor is updated based on a loss value between the extraction result obtained and a label of the clean speech data. The above process continues iteratively until a trained speech separation model is obtained. The acquiring of the clean sample set would be performed once for each training epoch, and in each training iteration of a training epoch, the inputting of the clean sample(s) into the speaker embedding extractor would be performed. For each training iteration, there would one or more clean samples used for training, which is not limited herein, and the following description would be made by taking one clean sample per training iteration as an example, but would also be applicable for the case of more than one clean sample per training iteration.

[0096] For example, the clean speech data used in different iterations would be different speech data of the training speaker, and the extraction result in each iteration would be an embedding of the corresponding clean speech data. The speaker embedding extractor processes the clean speech data to generate an extraction result, which is a representation or embedding vector that captures the distinctive characteristics of the training speaker's voice. A loss value is calculated based on the discrepancy

between the extraction result corresponding to the clean sample used in each training iteration and the label (or ground truth representation) of the clean speech data. At least one parameter of the speaker embedding extractor is updated based on the loss value. The update is done in a manner that minimizes the loss, thereby improving the accuracy of the embedding extraction over time. Notably, the loss value
5 used for updating the parameters does not include a speaker classification loss. This process of inputting clean samples, obtaining extraction results, calculating loss, and updating parameters continues iteratively until a speaker embedding extractor is trained to a satisfactory level of performance, as measured by its ability to accurately extract speaker embeddings from clean speech data.

10 [0097] By excluding the speaker classification loss, the model is optimized specifically for the task of speech separation without considering the effect of speaker classification. In addition, the model's performance in speech separation can be improved by focusing on embedding quality and generalization.

[0098] In a possible implementation, the loss value may be a scale-invariant signal-to-distortion ratio
15 (SI-SDR) value or a convolutive transfer function invariant signal-to-distortion ratio (CI-SDR) value. In this way, the model training can focus on optimizing the speaker embedding extractor to produce embeddings that are most useful for the separation task, as measured by SI-SDR or CI-SDR, rather than for speaker classification. In addition, the model training can be directed towards improving the quality of the separated speech.

20 [0099] FIG. 7 shows a schematic flowchart of a speech processing method in meeting system according to one or more example embodiments of the present disclosure. Illustratively, the speech processing method of the present disclosure can be implemented by a speech processing device provided by the present disclosure, and the speech processing device can be implemented by any software and/or hardware. For example, the speech processing device can be: a tablet computer, a
25 mobile phone (such as a folding screen mobile phone, a large screen mobile phone, etc.), a wearable device, a vehicle-mounted device, an augmented reality (AR)/virtual reality (VR) equipment, a laptop computer, a ultra-mobile personal computer (UMPC), a netbook, a personal digital assistant (PDA), a smart TV, a smart screen, a HD TV, a 4K TV, a smart speaker, a smart projector and other Internet of Things (IOT) devices. The specific types of electronic devices are not limited in the embodiments of
30 the present disclosure. In a possible implementation, the method can be applied in an application

scenario, in which in a meeting, multiple users A, B and C can use the same voice input device such as microphone. The voice input device may transmit acquired to-be-separated speech signal to a processing device, and the processing device may separate and distinguish a target speech signal of a target speaker from the to-be-separated speech signal. As shown in FIG. 7, the method can include the following steps.

[0100] S701, obtain a to-be-separated speech signal collected in a meeting system, where the to-be-separated speech signal includes target speech data of a target meeting participant and non-target speech data of at least one non-target meeting participant.

[0101] S702, process the to-be-separated speech signal with a pre-trained speech separation model based on a reference embedding corresponding to the target meeting participant, to obtain the target speech data of the target meeting participant, where the pre-trained speech separation model includes a separator containing a plurality of processing layers, and the plurality of processing layers include a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the reference embedding corresponding to the target meeting participant, a fused global embedding from the previous processing layer and an intermediate separation embedding, where the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

[0102] In the embodiment, the to-be-separated speech signal in a meeting system is obtained. The to-be-separated speech signal includes the target speech data of a specific meeting participant (whose speech needs to be extracted clearly) and non-target speech data from all other participants in the meeting. The processing device utilizes a pre-trained speech separation model that has been trained to process the captured speech signal. The speech separation model includes multiple processing layers, each designed to refine the separation of the target speech from the mixture. Each processing layer outputs an embedding that is a result of combining: a fused global embedding, which is a time-independent representation that captures the global features of the speech signal; and an output embedding from the preceding processing layer, which provides the context from the previous stage of separation. The fused global embedding is further refined using a reference embedding that corresponds to the target meeting participant (which may be obtained from a speaker embedding

extractor), a fused global embedding from the previous processing layer and an intermediate separation embedding that is derived from the processing layer for the input from the preceding layer. The processing and refinement of embeddings continue through each layer of the speech separation model until the final output is achieved, which ideally includes the clearly separated target speech data.

5 [0103] In a meeting system, the method provided in the embodiments of the present disclosure could be employed in real-time to isolate and enhance the speech of the target speaker, for example, a speaker who is speaking currently, while attenuating the background speech, thus improving the listening experience for remote participants. By applying the speech processing method within a conference system, it is possible to significantly improve the speech extraction performance, which is particularly
10 useful in noisy environments or when there is a need to focus on a specific speaker's contribution amidst multiple concurrent speakers.

[0104] By focusing on extracting the speech of a target participant, the method can significantly reduce background noise and crosstalk, leading to clearer and more intelligible speech. With better speech separation, the target speech data of the target meeting participant can be effectively obtained,
15 immediate and clear communication can be achieved among meeting participants.

[0105] FIG. 8 is a schematic illustration of a speech processing method according to one or more example embodiments of the present disclosure. As shown in FIG. 8, a speech separation model includes an encoder, a decoder and a separator containing multiple processing layers, the separator is shown as including multiple Single-Path Global Modulation (SPGM) Blocks (which is a specific
20 example of the above separation part of a processing layer or a specific example of the Sep Block). A to-be-separated speech signal is obtained through an encoder, and the to-be-separated speech signal is a mixture of a target speech signal of a target speaker and at least one non-target speech signal of at least one non-target speaker. Each of the multiple processing layers also includes a Global-Aware Auxiliary Fusion (GAAF) module (which is a specific example of the above fusion part of a processing layer),
25 the GAAF method is applied to the GAAF modules and the SPGM blocks to process the to-be-separated speech signal, and the GAAF method is the speech processing method according to one or more embodiments of the present disclosure. The to-be-separated speech signal is processed by the multiple SPGM Blocks and GAAF modules with the GAAF method. A target speech signal of the target speaker can be obtained through the decoder based on an output embedding output by a last
30 SPGM Block among the multiple SPGM Blocks and a speech encoding embedding of the

to-be-separated speech signal. It should be noted that although SPGM is shown illustratively, other SSMs are also possible, which is not limited in the embodiments of the present disclosure.

[0106] The GAAF method provides bidirectional interaction between SSMs and the speaker embedding extractor. Unlike fusion methods in related art, the speaker embedding provided to the SSM changes depending on the feature sequence. The GAAF module allows the identification task to be solved separately at a time-independent level of representation, and for the SSM to specialize on the local separation task it was designed for.

[0107] FIG. 9 is a schematic illustration of a speech processing method according to one or more example embodiments of the present disclosure. As shown in FIG. 9, a detailed view of the GAAF mechanism is illustrated.

[0108] As shown in FIG. 9, an overall model includes three main modules: an SSM (a part of which is shown as SPGM Block in the figure) which solves a separation problem using auxiliary information and delivers the model output, a speaker embedding extractor which extracts relevant features from the reference speech, and a GAAF fusion block (shown as GAAF in the figure) which allows bidirectional interaction between the primary SSM and the speaker embedding extractor. A systematic description of the flow of information through a TSE model shown in FIG. 9 is as follows:

$$\begin{aligned}
 x_0 &= \text{Encoder}(m_{in}) \\
 x_1, g_1 &= \text{SPGMBlock}_1(x_0, \text{GAAF}(a)) \\
 x_i, g_i &= \text{SPGMBlock}_i(x_{i-1}, \text{GAAF}(g_{x_{i-1}}, a)), i \in [2, N] \\
 T_{out} &= \text{Decoder}(\text{Mask}(x_N) \otimes x_0)
 \end{aligned} \tag{1}$$

where x_i represents the primary feature stream for the mixture (to-be-separated signal), g_i represents a global representation (or referred to as global embedding) of x_i produced by the SPGM Block and a represents a vector (or referred to as embedding) obtained from the speaker embedding extractor.

[0109] The GAAF mechanism facilitates a bidirectional interaction between the SSM's feature stream and the auxiliary embedding at the global level, shown in FIG. 9. The GAAF creates adaptability in the fusion of auxiliary speaker embeddings in two ways, (1) it adapts to a global representation of the SSM's intermediate embeddings and (2) it adapts to the global representation of the previous block of the SSM.

[0110] For (1), the global extraction (GlobExt, which is a specific example of the above mentioned

global extracting sub-layer) utilizes the last element selection proposed by SPGM to obtain a global embedding, g_{xi} , from the intermediate output of the SSM, x'_i . This is concatenated with the previous GAAF output and the auxiliary embedding from the speaker extractor, achieving (2). The Audio Mixer is implemented as a single linear layer that maps the concatenated dimension back to the original dimension of the SSM, as shown below:

$$\begin{aligned} g'_{x_i} &= \text{GlobExt}(x'_i) \\ r_i &= \text{Concat}(g'_{x_i}, g_{x_{i-1}}, a), r_i \in \mathbb{R}^{2G+A} \\ e_i &= \text{Linear}(r_i), r_i \in \mathbb{R}^G \end{aligned} \quad (2)$$

where G and A are the representational dimensions of the SSM and the speaker extractor respectively, g_{xi} is passed to the next GAAF fusion block and e_i is passed to a modulation layer in the SSM.

[0111] In this case the modulation method is implemented, which can be thought of as an adaptive filter on the SSM feature stream. The GAAF fusion block is attached to every SPGM Block, and creates a continuous global representation that is refined with each layer. The bidirectional interaction is created using the linear layer in the Audio Mixer, which fully connects the global representation of the SSM's intermediate features to the auxiliary embedding, allowing it to extract only relevant details for the SSM in each pass. The Audio Mixer and the Emb below the Audio Mixer can be regarded as a specific example of a combination of the above mentioned adaption sub-layer and fusion sub-layer.

[0112] For the primary SSM, an SPGM which offers good speech separation performance, while having convenient structural features for TSE. A last-element modulation method can be used to determine a global embedding in the current SPGM block. The last-element modulation method can take advantage of the innate redundancy of a dual path chunking method to guide a transformer layer into using the last element of a chunk as a global embedding token containing time-independent features. This helps extract good global features for GAAF, which is inserted between the global pooling and modulation layers of the SPGM block.

[0113] As described above, in many TSE models in the related art, speaker embedding extractors produce an embedding that is jointly trained on both a speaker classification task and on the TSE task. The goal of the secondary speaker classification loss is to guide the model towards more discriminative embeddings. The speaker classifier is a linear layer over the speaker embeddings producing speaker classes. The loss is computed using Cross Entropy (L_{CE}) and added to the main SI-SDR loss. However,

L_{CE} may be unnecessary because incorrect speaker classification also penalizes the main SI-SDR loss, and having a single loss prevents potential gradient conflicts. The exclusion of the speech classification loss may simplify the training objective, which can lead to more efficient and stable optimization of the model.

5 [0114] Unlike SEF-Net, which eliminates the auxiliary speaker extraction model along with the classification continues to rely on a dedicated speaker embedding extractor, using simply primary SI-SDR loss as the objective. Speaker embeddings can still be learnt without a classification loss because SI-SDR will still implicitly penalize the misclassification of speakers. Nevertheless, this creates some performance trade-offs.

10 [0115] The rationale for using a separate speaker embedding extractor instead of integrating the auxiliary information into the SSM, like in SEF-Net, is that GAAF relies on modelling of global and local features at different representational resolutions. As the work in SPGM showed, modelling local features should be assigned a greater number of parameters while global, time-independent, features can be modelled with a coarser representation and fewer parameters.

15 [0116] The performance of the above SSM can be tested by several experiments. The experiments and results will be described in details.

[0117] **Dataset**

[0118] A WSJ0-2mix-extr dataset was generated. This dataset is chosen because it offers maximum backward compatibility of benchmark against many recent and prior models. The WSJ0-2mix-extr
20 dataset includes 101 speakers in the training and development set, with 20,000 and 5,000 mixture utterances respectively. The test set includes an additional 18 speakers, unseen in the training and development sets, across 3,000 mixture utterances. A first speaker of each mixture is used as a target speaker and a random utterance from all available utterances from that speaker (excluding the one already in the mixture) is selected to be the reference.

25 [0119] **Training Parameters**

[0120] All models were trained using the Speechbrain framework. We train all models for 150 epochs without any data augmentations. The Adam optimizer was used with a starting learning rate of $1.5e^{-4}$ and a weight decay of 0. The learning rate is kept constant until epoch 85 and is then halved with a patience of 2. SI-SDR is used as the loss function. Where classification loss is required, cross entropy
30 is used.

[0121] Model Parameters

[0122] The headline model as reported in Table 1 uses $N = 4$ SPGM Blocks, an SSM embedding dimension (G) of 256, an auxiliary speaker embedding dimension (A) of 1024, and no cross entropy loss was used during training. For some of the ablation studies, a speaker embedding dimension (A) of 512 was used for efficiency, which is denoted by GAAF-512 in Table 3 and Table 4.

[0123] The performance of GAAF is compared against other baseline models in Table 1. Backbone refers to the SSM architecture used in the TSE model. The GAAF model achieved the best performance among all the models in terms of SDR, SDRi, SISDR, and SI-SDRi metrics. It also achieved a competitive PESQ score, with only X-Sepformer performing marginally better. Additionally, the SI-SDRi of GAAF is 19.4dB, which outperforms X-Sepformer (S_{base}) by 0.5dB. Notably, these results are achieved without using speaker classification loss.

[0124] Compared against the most recent model that also adopts this approach, SEF-Net, GAAF surpasses SEF-Net on all metrics and by a margin of 2.1 in SDRi and 2.2 in SI-SDRi, demonstrating the effectiveness of the proposed architecture even without relying on speaker classification. This suggests that GAAF captures speaker-discriminative features directly during the separation process, without the need for an explicit speaker classification loss.

Table 1: Performance of GAAF with respect to other speaker extraction models

Model	Backbone	Speaker Classification	Performance				
			SDR	SDRi	SI-SDR	SI-SDRi	PESQ
TseNet	ConvTasNet	✓	15.2	-	14.7	-	3.14
SpEx	ConvTasNet	✓	17.2	17.0	16.7	16.6	3.36
SpEx+	ConvTasNet	✓	18.5	17.6	18.2	17.4	3.49
WASE	ConvTasNet	✓	17.2	17.1	-	-	-
SpEx _{pc}	ConvTasNet	✓	-	19.2	-	19.0	-
X-Sepformer(S_{base})	Sepformer	✓	-	19.5	-	18.9	3.74
SEF-Net	Conformer	×	19.3	17.6	18.8	17.2	3.50
GAAF (Ours)	SPGM	×	19.9	19.7	19.4	19.4	3.71

[0125] While X-Sepformer (S_{base}) achieves the highest PESQ score, GAAF remains competitive, indicating good perceptual quality alongside its superior separation metrics. Overall, these results highlight the strong performance of GAAF in speaker extraction, particularly its ability to achieve state-of-the-art results without relying on an explicit speaker classification step. It can be seen that the speaker classification loss may be not necessary for the TSE task.

Table 2: Ablation Study of GAAF with various auxiliary embedding sizes and use of previous global embedding

Emb Size	Prev Glob	Performance			
		SDR	SDRi	SI-SDR	SI-SDRi
256	×	18.3	18.1	17.2	17.2
512	×	19.6	19.5	19.0	19.0
1024	×	19.6	19.4	19.0	19.0
512	✓	19.8	19.6	19.3	19.3
1024	✓	19.9	19.7	19.4	19.4

[0126] Ablation on Speaker Embedding Sizes

[0127] To show that the auxiliary speaker embedding features are indeed helpful to the performance of the model despite going through the GAAF fusion block, an ablation study was performed across multiple auxiliary speaker embedding sizes (A) and the results are shown in Table 2. Two sets of experiments were performed, each depending on whether we make use of the output of the previous GAAF block. The representation size of the underlying SSM model (G) is kept constant at 256. Our results show that as the size of the speaker embedding increases, the performance of the model increases. However, beyond a size of 512, there is no improvement in performance without including an additional path into GAAF utilizing the previous global (Prev Glob) embedding. Additionally, the experiments where the previous global embedding is used always outperform experiments where they are not used. Overall, these results demonstrate the importance of the main features of GAAF.

Table 3: Performance of GAAF based on pre-trained auxiliary embedding extractors

Model	Performance			
	SDR	SDRi	SI-SDR	SI-SDRi
GAAF-512	19.8	19.6	19.3	19.3
+ pre-trained	19.7	19.5	19.2	19.2
+ frozen	17.1	16.9	14.7	14.7

[0128] Impact of pre-training auxiliary encoder

[0129] One common workflow for TSE models is the pre-training of the auxiliary speaker extractor, which we implement here and report the results in Table 3. We pre-train the auxiliary encoder as described in the original ACA-Net paper and achieve an EER of 2.34 and minDCF of 0.28 on the WSJ0-1talker training dataset which shares the same set of speakers in the training set as our

WSJ0-2mix-extr training set. This is intended to improve speaker embedding quality by allowing the auxiliary model to pre-train on the speaker verification task. However, as our results in Table 3 show, when the model is pre-trained but not fine-tuned, the model achieves similar performance to if it was randomly initialized. Additionally, if the weights of the speaker extractor are frozen, the model performs poorly. This result suggests that the embeddings extracted for the speaker verification task are not always useful for TSE, since the frozen pre-trained model performs poorly. Even with GAAF acting as an adapter layer for the frozen layer, the limited parameters of the GAAF mixer layer were not able to adapt the frozen speaker embeddings enough to deliver good performance.

[0130] Impact of removing L_{CE}

[0131] In this work, the speaker classifier and the cross entropy loss were removed in the training of GAAF based on the insight from SEF-NET that the speaker classification loss may be unnecessary. Unlike SEF-Net, the GAAF model of the present maintains a separate auxiliary speaker extraction model. Based on the results in Table 4 it can be seen that removing L_{CE} does indeed lead to performance improvement for GAAF.

[0132] One of the main issues impacting TSE performance is the speaker confusion problem, which can be visualized using a histogram of the SI-SDRi scores over each sample in the test set. The histogram is plotted with the models reported in FIG. 10. FIG. 10 is a schematic diagram of a histogram of GAAF with (CE) and without (noCE) Cross-Entropy loss. As shown in FIG. 10, the histogram of the model trained with CE has three peaks, one above 20dB SI-SDRi (correct target, good separation), one just above 0dB SI-SDRi (correct target, bad separation), and one at -40dB SI-SDRi (incorrect target). For the histogram of the model trained without using CE, it can be seen that the -40dB peak is mostly absent, with most of the samples moved to the 0dB and 20dB peaks.

[0133] When computing the average of only samples with scores >10 dB SI-SDRi, the model without L_{CE} achieves 19.5dB while the model with L_{CE} achieves 21.5dB, comparable to the backbone model's SS performance of 22.1 dB. This shows that TSE models are competitive with SS models, but the overall result reported in most studies is pulled down by the speaker confused samples in the test set. It may be because the model might over fit on the speakers in the training set when L_{CE} is used, resulting in more frequent misclassification on the test set. Removing CE eliminates this issue, although increases the number of cases with correct target and bad separation. It is due to chunk-level speaker confusion and can be remedied in future work using the loss functions. Finally, this histogram

also highlights the difficulty of directly comparing TSE and SSM results, given that TSE consists of two sub-tasks while SSMs only have a single task.

Table 4: Performance of GAAF with and without speaker classification loss

GAAF-512	Performance			
	SDR	SDRi	SI-SDR	SI-SDRi
with L_{CE}	16.5	16.4	14.1	14.1
without L_{CE}	19.6	19.5	19.0	19.0

[0134] According to the above experiments and the results, the GAAF of the present disclosure can deliver strong performance on TSE by facilitating bidirectional interaction between the auxiliary speaker embedding extractor and the primary SSM. The GAAF outperforms all recent strong baselines on this task and improves upon the performance of the strongest baseline by 0.6dB SDR. The performance of the model of the present disclosure can be attributed to several innovations of the GAAF mechanism, including the adaptation of the auxiliary speaker embedding to a global representation of the mixture using the embedding mixer, the transfer of global embeddings through the SSM layers, and the removal of speaker classification loss during the training of the model.

[0135] FIG. 11 shows a schematic structural diagram of a speech processing apparatus according to one or more embodiments of the present disclosure. As shown in FIG. 11, a speech processing apparatus 1100 may include:

15 an obtaining module 1101, configured to obtain a to-be-separated speech signal, where the to-be-separated speech signal is a mixture of a target speech signal of a target speaker and at least one non-target speech signal of at least one non-target speaker;

a processing module 1102, configured to process, based on a reference embedding corresponding to the target speaker, the to-be-separated speech signal with a pre-trained speech separation model to obtain the target speech signal of the target speaker, where the pre-trained speech separation model includes a separator containing a plurality of processing layers and the plurality of processing layers include a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the reference embedding corresponding to the target speaker, a fused global embedding from the previous processing layer and

an intermediate separation embedding, where the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

[0136] In a possible implementation, the speech separation model further includes an encoder and a decoder;

5 the processing module 1102 is configured to:

obtain, through the encoder, a speech encoding embedding based on the to-be-separated speech signal;

for an i -th processing layer among the plurality of processing layers of the separator, obtain an i -th intermediate separation embedding based on a first input of the i -th processing layer, generate an i -th fused global embedding based on a second input of the i -th processing layer and an i -th global embedding of the i -th intermediate separation embedding, and output an i -th output embedding based on the i -th intermediate separation embedding and the i -th fused global embedding; where i is a positive integer not greater than N , N is a number of the plurality of processing layers, when i equals to 1, the first input of the i -th processing layer is the speech encoding embedding, and the second input of the i -th processing layer is the reference embedding; when i is greater than 1 but not greater than N , the first input of the i -th processing layer is an $(i-1)$ -th output embedding from an $(i-1)$ -th processing layer, and the second input of the i -th processing layer is the reference embedding and an $(i-1)$ -th fused global embedding from the $(i-1)$ -th processing layer;

20 obtain, through the decoder, the target speech signal of the target speaker based on an output embedding output by an N -th processing layer among the plurality of processing layers and the speech encoding embedding.

[0137] In a possible implementation, a processing layer of the plurality of processing layers includes a separation part and a fusion part;

the processing module 1102 is configured to:

25 input the first input of the i -th processing layer into a separation part of the i -th processing layer to obtain the i -th intermediate separation embedding;

input the i -th intermediate separation embedding and the second input of the i -th processing layer into a fusion part of the i -th processing layer to obtain the i -th fused global embedding, and output the i -th fused global embedding to the separation part of the i -th processing layer;

30 output the i -th output embedding by the separation part of the i -th processing layer based on

the i-th intermediate separation embedding and the i-th fused global embedding from the fusion part of the i-th processing layer.

[0138] In a possible implementation, the fusion part of the i-th processing layer includes a global extracting sub-layer, an adaption sub-layer and a fusion sub-layer;

5 the processing module 1102 is configured to:

obtain, through the global extracting sub-layer, the i-th global embedding of the i-th intermediate separation embedding;

adapt, through the adaption sub-layer, the second input of the i-th processing layer and the i-th global embedding of the i-th intermediate separation embedding to processing at the fusion sub-layer; and

10

fuse, through the fusion sub-layer, the adapted second input of the i-th processing layer and the i-th adapted global embedding of the i-th intermediate separation embedding to obtain the i-th fused global embedding, where a size of the i-th fused global embedding is matched with a size of the i-th global embedding of the i-th intermediate separation embedding.

15 [0139] In a possible implementation, the processing module 1102 is configured to:

perform at least one of a rearrangement, filtering or enhancing operation on elements of the i-th global embedding of the i-th intermediate separation embedding and each embedding in the second input of the i-th processing layer to obtain the adapted second input of the i-th processing layer and the i-th adapted global embedding of the i-th intermediate separation embedding.

20 [0140] In a possible implementation, the processing module 1102 is configured to:

input the i-th intermediate separation embedding into the global extracting sub-layer and determine the last element from the i-th intermediate separation embedding as the i-th global embedding.

[0141] In a possible implementation, the adaption sub-layer is a single convolutional network.

25 [0142] In a possible implementation, the processing module 1102 is configured to:

concatenate the i-th adapted global embedding of the i-th intermediate separation embedding and each embedding in the adapted second input of the i-th processing layer, and extract a number of weighted elements in the concatenated embedding as the i-th fused global embedding, where the number of weighted elements in the concatenated embedding is as same as a size of the i-th global embedding of the i-th intermediate separation embedding; or,

30

align and averaging the i-th adapted global embedding of the i-th intermediate separation embedding and each embedding in the second input of the i-th processing layer, and obtain the i-th fused global embedding based on an averaged embedding, where the averaged embedding is an average of the i-th adapted global embedding of the i-th intermediate separation embedding and each embedding in the second input of the i-th processing layer.

[0143] In this way, information from both the adapted global embedding and the adapted second input can be integrated, a more comprehensive representation that can capture a wider range of speaker characteristics can be generated. In addition, the size of the fused global embedding can be ensured to match the size of the original global embedding, which maintains consistency in the dimensionality of the embeddings throughout the processing layers.

[0144] In a possible implementation of the first aspect, the separation part of the i-th processing layer is a single-path global modulation (SPGM) block. The SPGM block can be designed to modulate the speech representation in a way that emphasizes the target speaker's features, which can lead to more effective separation.

[0145] In a possible implementation, the processing module 1102 is configured to:
process a reference speech signal with a pre-trained speaker embedding extractor to obtain the reference embedding corresponding to the target speaker, where the reference speech signal includes clean speech data of the target speaker, and the pre-trained speaker embedding extractor is obtained from a training for speech extraction during which a speech classification loss is skipped.

[0146] In a possible implementation, the pre-trained speaker embedding extractor is a pre-trained Asymmetric Cross Attention (ACA) net with a frequency-domain encoder.

[0147] FIG. 12 shows a schematic structural diagram of a model training apparatus according to one or more embodiments of the present disclosure. As shown in FIG. 12, a model training apparatus 1200 may include:

a processing module 1201, configured to:
acquire a training sample set, wherein the training sample set comprises a plurality of training samples and a sample reference embedding corresponding to a training speaker among a plurality of speakers, and each of the plurality of training samples is a mixture of speech signals from the plurality of speakers;

for a training sample in the training sample set, input the training sample and the sample

reference embedding into a speech separation model to obtain a separation result corresponding to the training sample;

an updating module 1202, configured to update at least one parameter of the speech separation model based on the separation result;

5 where the speech separation model comprises a separator containing a plurality of processing layers and the plurality of processing layers include a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the sample reference embedding corresponding to the training speaker, a fused global embedding from the previous processing layer and an intermediate separation embedding, where the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

10 [0148] FIG. 13 shows a schematic structural diagram of a model training apparatus according to one or more embodiments of the present disclosure. As shown in FIG. 13, a model training apparatus 1300 may include:

 a processing module 1301, configured to:

 acquire a clean sample set including a plurality of clean samples, where each of the plurality of clean samples includes clean speech data of a training speaker;

20 for a clean sample in the clean sample set, input a clean sample into a speaker embedding extractor to obtain an extraction result;

 an updating module 1302, configured to update at least one parameter of the speaker embedding extractor based on a loss value between the extraction result and a label of the clean speech data, where a speaker classification loss is excluded from the loss value.

25 [0149] In a possible implementation, the loss value is a scale-invariant signal-to-distortion ratio (SI-SDR) value or a convolutive transfer function invariant signal-to-distortion ratio (CI-SDR) value.

 [0150] It should be understood by a person skilled in the art that, the relevant description of the above modules in the possible implementations of the present disclosure may be understood with reference to the relevant description of the speech separation method in the possible implementations of the present disclosure. The technical effects achieved by the above apparatuses are similar as those

achieved by the above corresponding method embodiments, which is not repeated herein.

[0151] FIG. 14 is a structural diagram of an electronic device according to one or more embodiments of the present disclosure. As shown in FIG. 14, the electronic device 1400 may include: a processor 1401 coupled with a memory 1402 in a communicative way via an interface 1403; where the memory 1402 stores a computer executable instruction; the processor 1401 executes the computer executable instruction stored in the memory 1402 for executing any of the above speech processing methods. It should be noted that, the memory 1402 may be included or excluded from the electronic device, depending on actual needs.

[0152] FIG. 15 is a structural diagram of another electronic device according to one or more embodiments of the present disclosure. As shown in FIG. 15, the electronic device 1500 may include: a processor 1501 coupled with a memory 1502 in a communicative way via an interface 1503; where the memory 1502 stores a computer executable instruction; the processor 1501 executes the computer executable instruction stored in the memory 1502 for executing any of the above model training methods. It should be noted that, the memory 1502 may be included or excluded from the electronic device, depending on actual needs.

[0153] An embodiment of the present disclosure provides a speech processing apparatus, including:

an obtaining module configured to obtain a to-be-separated speech signal collected in a meeting system, where the to-be-separated speech signal includes target speech data of a target meeting participant and non-target speech data of at least one non-target meeting participant;

a processing module configured to process, based on a reference embedding corresponding to the target meeting participant, the to-be-separated speech signal with a pre-trained speech separation model to obtain the target speech data of the target meeting participant, where the pre-trained speech separation model includes a separator containing a plurality of processing layers and the plurality of processing layers include a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the reference embedding corresponding to the target meeting participant, a fused global embedding from the previous processing layer and an intermediate separation embedding, where the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

[0154] An embodiment of the present disclosure provides a computer-readable medium storing computer execution instructions which, when executed by a processor, causes the processor to execute any of the above methods.

5 [0155] An embodiment of the present disclosure provides a computer program product including computer execution instructions which, when executed by a processor, causes the processor to execute any of the above methods.

[0156] An embodiment of the present disclosure provides a computer program which, when executed by a processor, causes the processor to execute any of the above methods.

10 [0157] The present disclosure encompasses various embodiments, including not only method embodiments, but also other embodiments such as apparatus embodiments and embodiments related to non-transitory computer readable storage media. Embodiments may incorporate, individually or in combinations, the features disclosed herein.

[0158] Although this disclosure refers to illustrative embodiments, this is not intended to be construed in a limiting sense. Various modifications and combinations of the illustrative embodiments, 15 as well as other embodiments of the disclosure, will be apparent to persons skilled in the art upon reference to the description.

[0159] Features disclosed herein in the context of any particular embodiments may also or instead be implemented in other embodiments. Method embodiments, for example, may also or instead be implemented in apparatus, system, and/or computer program product embodiments. In addition, 20 although embodiments are described primarily in the context of methods and apparatus, other implementations are also contemplated, as instructions stored on one or more non-transitory computer-readable media, for example. Such media could store programming or instructions to perform any of various methods consistent with the present disclosure.

[0160] Although the present disclosure describes methods and processes with steps in a certain order, 25 one or more steps of the methods and processes may be omitted or altered as appropriate. One or more steps may take place in an order other than that in which they are described, as appropriate.

[0161] Note that the expression “at least one of A or B”, as used herein, is interchangeable with the expression “A and/or B”. It refers to a list in which you may select A or B or both A and B. Similarly, “at least one of A, B, or C”, as used herein, is interchangeable with “A and/or B and/or C” or “A, B, 30 and/or C”. It refers to a list in which you may select: A or B or C, or both A and B, or both A and C, or

both B and C, or all of A, B and C. The same principle applies for longer lists having a same format.

[0162] Although the present disclosure is described, at least in part, in terms of methods, a person of ordinary skill in the art will understand that the present disclosure is also directed to the various components for performing at least some of the aspects and features of the described methods, be it by way of hardware components, software or any combination of the two. Accordingly, the technical solution of the present disclosure may be embodied in the form of a software product. A suitable software product may be stored in a pre-recorded storage device or other similar non-volatile or non-transitory computer readable medium, including DVDs, CD-ROMs, USB flash disk, a removable hard disk, or other storage media, for example. The software product includes instructions tangibly stored thereon that enable a processing device (e.g., a personal computer, a server, or a network device) to execute examples of the methods disclosed herein. The machine-executable instructions may be in the form of code sequences, configuration information, or other data, which, when executed, cause a machine (e.g., a processor or other processing device) to perform steps in a method according to examples of the present disclosure.

[0163] The present disclosure may be embodied in other specific forms without departing from the subject matter of the claims. The described example embodiments are to be considered in all respects as being only illustrative and not restrictive. Selected features from one or more of the above-described embodiments may be combined to create alternative embodiments not explicitly described, features suitable for such combinations being understood within the scope of this disclosure.

[0164] All values and sub-ranges within disclosed ranges are also disclosed. Also, although the systems, devices and processes disclosed and shown herein may include a specific number of elements/components, the systems, devices and assemblies could be modified to include additional or fewer of such elements/components. For example, although any of the elements/components disclosed may be referenced as being singular, the embodiments disclosed herein could be modified to include a plurality of such elements/components. The subject matter described herein intends to cover and embrace all suitable changes in technology.

[0165] Although embodiments have been described above with reference to the accompanying drawings, those of skill in the art will appreciate that variations and modifications may be made without departing from the scope thereof as defined by the appended claims.

CLAIMS

1. A speech processing method, comprising:

obtaining a to-be-separated speech signal, wherein the to-be-separated speech signal is a mixture of a target speech signal of a target speaker and at least one non-target speech signal of at least one non-target speaker;

processing, based on a reference embedding corresponding to the target speaker, the to-be-separated speech signal with a pre-trained speech separation model to obtain the target speech signal of the target speaker, wherein the pre-trained speech separation model comprises a separator containing a plurality of processing layers, and the plurality of processing layers comprise a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the reference embedding corresponding to the target speaker, a fused global embedding from the previous processing layer and an intermediate separation embedding, wherein the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

2. The method according to claim 1, wherein the speech separation model further comprises an encoder and a decoder;

wherein the processing, based on the reference embedding corresponding to the target speaker, the to-be-separated speech signal with the pre-trained speech separation model to obtain the target speech signal of the target speaker comprises:

obtaining, through the encoder, a speech encoding embedding based on the to-be-separated speech signal;

for an i-th processing layer among the plurality of processing layers of the separator, obtaining an i-th intermediate separation embedding based on a first input of the i-th processing layer, generating an i-th fused global embedding based on a second input of the i-th processing layer and an i-th global embedding of the i-th intermediate separation embedding, and outputting an i-th output embedding based on the i-th intermediate separation embedding and the i-th fused global embedding; wherein i is a positive integer not greater than N, N is a number of the plurality of processing layers, when i equals to 1, the first input of the

i-th processing layer is the speech encoding embedding, and the second input of the i-th processing layer is the reference embedding; when i is greater than 1 but not greater than N, the first input of the i-th processing layer is an (i-1)-th output embedding from an (i-1)-th processing layer, and the second input of the i-th processing layer is the reference embedding and an (i-1)-th fused global embedding from the (i-1)-th processing layer;

obtaining, through the decoder, the target speech signal of the target speaker based on an output embedding output by an N-th processing layer among the plurality of processing layers and the speech encoding embedding.

3. The method according to claim 2, wherein a processing layer of the plurality of processing layers comprises a separation part and a fusion part;

wherein the obtaining an i-th intermediate separation embedding based on the first input of the i-th processing layer comprises:

inputting the first input of the i-th processing layer into a separation part of the i-th processing layer to obtain the i-th intermediate separation embedding;

wherein generating an i-th fused global embedding based on the second input of the i-th processing layer and an i-th global embedding of the i-th intermediate separation embedding comprises:

inputting the i-th intermediate separation embedding and the second input of the i-th processing layer into a fusion part of the i-th processing layer to obtain the i-th fused global embedding, and outputting the i-th fused global embedding to the separation part of the i-th processing layer;

wherein outputting an i-th output embedding based on the i-th intermediate separation embedding and the i-th fused global embedding comprises:

outputting the i-th output embedding by the separation part of the i-th processing layer based on the i-th intermediate separation embedding and the i-th fused global embedding from the fusion part of the i-th processing layer.

4. The method according to claim 3, wherein the fusion part of the i-th processing layer comprises a global extracting sub-layer, an adaption sub-layer and a fusion sub-layer;

wherein the inputting the i-th intermediate separation embedding and the second input of the i-th processing layer into the fusion part of the i-th processing layer to obtain the i-th fused global embedding comprises:

obtaining, through the global extracting sub-layer, the i-th global embedding of the i-th intermediate

separation embedding;

adapting, through the adaption sub-layer, the second input of the i-th processing layer and the i-th global embedding of the i-th intermediate separation embedding to processing at the fusion sub-layer; and

5 fusing, through the fusion sub-layer, the adapted second input of the i-th processing layer and the i-th adapted global embedding of the i-th intermediate separation embedding to obtain the i-th fused global embedding, wherein a size of the i-th fused global embedding is matched with a size of the i-th global embedding of the i-th intermediate separation embedding .

5. The method according to claim 4, wherein the adaption of the second input of the i-th processing layer and the i-th global embedding of the i-th intermediate separation embedding comprises:

10 performing at least one of a rearrangement, filtering or enhancing operation on elements of the i-th global embedding of the i-th intermediate separation embedding and each embedding in the second input of the i-th processing layer to obtain the adapted second input of the i-th processing layer and the i-th adapted global embedding of the i-th intermediate separation embedding.

6. The method according to claim 4 or 5, wherein the obtaining, through the global extracting sub-layer, the i-th global embedding of the i-th intermediate separation embedding comprises:

inputting the i-th intermediate separation embedding into the global extracting sub-layer and determining the last element from the i-th intermediate separation embedding as the i-th global embedding.

7. The method according to any one of claims 4 to 6, wherein the adaption sub-layer is a single convolutional network.

20 8. The method according to any one of claims 3 to 7, wherein the fusing of the adapted second input of the i-th processing layer and the i-th adapted global embedding of the i-th intermediate separation embedding to obtain the i-th fused global embedding comprises:

concatenating the i-th adapted global embedding of the i-th intermediate separation embedding and each embedding in the adapted second input of the i-th processing layer, and extracting a number of weighted elements in the concatenated embedding as the i-th fused global embedding, wherein the number of weighted elements in the concatenated embedding is as same as a size of the i-th global embedding of the i-th intermediate separation embedding; or,

aligning and averaging the i-th adapted global embedding of the i-th intermediate separation embedding and each embedding in the second input of the i-th processing layer, and obtaining the i-th fused global embedding based on an averaged embedding, wherein the averaged embedding is an average of the i-th

adapted global embedding of the i-th intermediate separation embedding and each embedding in the second input of the i-th processing layer.

9. The method according to any one of claims 3 to 8, wherein the separation part of the i-th processing layer is a single-path global modulation (SPGM) block.

5 10. The method according to any one of claims 1 to 9, further comprising:

processing a reference speech signal with a pre-trained speaker embedding extractor to obtain the reference embedding corresponding to the target speaker, wherein the reference speech signal comprises clean speech data of the target speaker, and the pre-trained speaker embedding extractor is obtained from a training for speech extraction during which a speech classification loss is skipped.

10 11. The method according to claim 10, wherein the pre-trained speaker embedding extractor is a pre-trained Asymmetric Cross Attention (ACA) net with a frequency-domain encoder.

12. A model training method, comprising:

acquiring a training sample set, wherein the training sample set comprises a plurality of training samples and a sample reference embedding corresponding to a training speaker among a plurality of speakers, and
15 each of the plurality of training samples is a mixture of speech signals from the plurality of speakers;

for a training sample in the training sample set, inputting the training sample and the sample reference embedding into a speech separation model to obtain a separation result corresponding to the training sample, and updating at least one parameter of the speech separation model based on the separation result;

wherein the speech separation model comprises a separator containing a plurality of processing layers,
20 and the plurality of processing layers comprise a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the sample reference embedding corresponding to the training speaker, a fused global embedding from the previous processing layer and an
25 intermediate separation embedding, wherein the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

13. A speech processing method, comprising:

obtaining a to-be-separated speech signal collected in a meeting system, wherein the to-be-separated speech signal comprises target speech data of a target meeting participant and non-target speech data of at
30 least one non-target meeting participant;

processing, based on a reference embedding corresponding to the target meeting participant, the to-be-separated speech signal with a pre-trained speech separation model to obtain the target speech data of the target meeting participant, wherein the pre-trained speech separation model comprises a separator containing a plurality of processing layers, and the plurality of processing layers comprise a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the reference embedding corresponding to the target meeting participant, a fused global embedding from the previous processing layer and an intermediate separation embedding, wherein the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

14. A speech processing apparatus, comprising:

an obtaining module, configured to obtain a to-be-separated speech signal, wherein the to-be-separated speech signal is a mixture of a target speech signal of a target speaker and at least one non-target speech signal of at least one non-target speaker;

a processing module, configured to process, based on a reference embedding corresponding to the target speaker, the to-be-separated speech signal with a pre-trained speech separation model to obtain the target speech signal of the target speaker, wherein the pre-trained speech separation model comprises a separator containing a plurality of processing layers; wherein the plurality of processing layers comprise a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the reference embedding corresponding to the target speaker, a fused global embedding from the previous processing layer and an intermediate separation embedding, wherein the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

15. A model processing apparatus, comprising:

a processing module, configured to:

acquire a training sample set, wherein the training sample set comprises a plurality of training samples and a sample reference embedding corresponding to a training speaker among a plurality of speakers,

and each of the plurality of training samples is a mixture of speech signals from the plurality of speakers;

for a training sample in the training sample set, input the training sample and the sample reference embedding into a speech separation model to obtain a separation result corresponding to the training sample;

an updating module, configured to update at least one parameter of the speech separation model based on the separation result;

wherein the speech separation model comprises a separator containing a plurality of processing layers, and the plurality of processing layers comprise a first processing layer and remaining processing layers; for a processing layer of the remaining processing layers, the processing layer is configured to output an output embedding based on a fused global embedding and an output embedding from a previous processing layer of the processing layer, and the fused global embedding is refined with the sample reference embedding corresponding to the training speaker, a fused global embedding from the previous processing layer and an intermediate separation embedding, wherein the intermediate separation embedding is obtained by the processing layer based on the output embedding from the previous processing layer.

16. An electronic device, comprising at least one processor coupled with a memory storing a set of instructions;

wherein the at least one processor is configured to read the set of instructions in the memory and execute the speech processing method according to any one of claims 1 to 11, or the model training method according to claim 12, or the speech processing method according to claim 13.

17. A computer-readable medium storing computer execution instructions which, when executed by a processor, cause the processor to execute the speech processing method according to any one of claims 1 to 11, or the model training method according to claim 12, or the speech processing method according to claim 13.