WHAT IS CLAIMED IS:

1. A method for isolating a sound, comprising:

acquiring input sound spectra, the input sound spectra comprising sound spectra corresponding to multiple sound sources;

isolating predicted sound spectra from the input sound spectra by performing spectrum isolation processing on the input sound spectra;

acquiring updated input sound spectra by removing the predicted sound spectra from the input sound spectra; and

continuing to acquire next isolated predicted sound spectra through the updated input sound spectra, until the updated input sound spectra comprise no sound spectrum.

2. The method of claim 1, wherein isolating the predicted sound spectra from the input sound spectra by performing spectrum isolation processing on the input sound spectra comprises:

acquiring an input video frame corresponding to the input sound spectra, the input video frame comprising the multiple sound sources; and

isolating the predicted sound spectra from the input sound spectra by performing spectrum isolation processing on the input sound spectra according to the input video frame.

3. The method of claim 2, wherein isolating the predicted sound spectra from the input sound spectra by performing spectrum isolation processing on the input sound spectra according to the input video frame comprises:

acquiring k basic components according to the input sound spectra, the k basic components representing respective sound features in the input sound spectra, the k being a natural number;

acquiring a visual feature map according to the input video frame, the visual feature map comprising multiple visual feature vectors in k dimensions, each visual feature vector of the multiple visual feature vectors corresponding to one sound source in the input video frame;

and

acquiring the predicted sound spectra according to a visual feature vector of the multiple visual feature vectors as well as the k basic components, a sound source of the predicted sound spectra being a sound source corresponding to the visual feature vector.

4. The method of claim 3, wherein acquiring the visual feature map according to the input video frame comprises:

outputting a video feature of the input video frame by inputting the input video frame to a feature extraction network; and

acquiring the visual feature map comprising the multiple visual feature vectors by performing max pooling on the video feature in time dimension.

5. The method of claim 3, wherein acquiring the predicted sound spectra according to the visual feature vector of the multiple visual feature vectors as well as the k basic components comprises:

acquiring the predicted sound spectra as a dot product of a vector of the k basic components and the visual feature vector of k elements.

6. The method of claim 3, wherein acquiring the predicted sound spectra according to the visual feature vector of the multiple visual feature vectors as well as the k basic components comprises:

acquiring a dot product of a vector of the k basic components and the visual feature vector of k elements;

acquiring a predicted mask by performing nonlinear activation processing on the dot product; and

acquiring the predicted sound spectra as a dot product of the predicted mask and initial input sound spectra for a first iteration.

7. The method of claim 3, wherein acquiring the predicted sound spectra according to the visual feature vector of the multiple visual feature vectors as well as the k basic components comprises:

randomly selecting a visual feature vector from the multiple visual feature vectors; and

acquiring the predicted sound spectra according to the visual feature vector selected

and the k basic components.

8. The method of claim 3, wherein acquiring the predicted sound spectra according to the visual feature vector of the multiple visual feature vectors as well as the k basic components comprises:

selecting, from the multiple visual feature vectors, a visual feature vector corresponding to a loudest sound source; and

acquiring the predicted sound spectra according to the visual feature vector selected and the k basic components.

9. The method of claim 8, wherein selecting the visual feature vector corresponding to the loudest sound source comprises:

acquiring a first dot product of a vector of the k basic components and the each visual feature vector of the multiple visual feature vectors;

acquiring a second dot product of the first dot product having been subject to nonlinear activation and initial input sound spectra for a first iteration;

acquiring average energy of the second dot product; and

selecting a visual feature vector corresponding to a location of max average energy.

10. The method of any one of claims 1 to 9, further comprising: after isolating the predicted sound spectra from the input sound spectra,

acquiring a residual mask according to the predicted sound spectra and historical cumulative spectra, the historical cumulative spectra being a sum of historical predicted sound spectra isolated during sound isolation;

acquiring residual spectra based on the residual mask and the historical cumulative spectra; and

acquiring complete predicted sound spectra as a sum of the residual spectra and the predicted sound spectra.

11. The method of claim 10, wherein the sum of the historical predicted sound spectra comprises a sum of historical complete predicted sound spectra;

wherein acquiring the updated input sound spectra by removing the predicted sound spectra from the input sound spectra comprises:

acquiring the updated input sound spectra by removing the complete predicted sound spectra from the input sound spectra.

12. The method of claim 10, further comprising: adjusting a network parameter of at least any one of a first network, a second network, and a third network according to an error between the complete predicted sound spectra and true spectra,

wherein k basic components are acquired by inputting the input sound spectra to the first network, wherein a visual feature map is acquired by inputting an input video frame corresponding to the input sound spectra to the second network, wherein the input video frame comprises the multiple sound sources, wherein the residual mask is acquired by inputting the predicted sound spectra and the historical cumulative spectra to the third network.

13. The method of any one of claims 1 to 12, further comprising:

in response to average energy of the updated input sound spectra being less than a preset threshold, determining that the input sound spectra comprise no sound spectrum.

14. A device for isolating a sound, comprising an input acquiring module, a spectrum isolating module, and a spectrum updating module,

wherein the input acquiring module is configured to acquire input sound spectra, the input sound spectra comprising sound spectra corresponding to multiple sound sources,

wherein the spectrum isolating module is configured to isolate predicted sound spectra from the input sound spectra by performing spectrum isolation processing on the input sound spectra; and continue to acquire next isolated predicted sound spectra through updated input sound spectra, until the updated input sound spectra comprise no sound spectrum,

wherein the spectrum updating module is configured to acquire the updated input sound spectra by removing the predicted sound spectra from the input sound spectra.

15. The device of claim 14, wherein the spectrum isolating module comprises a video

processing sub-module and a sound isolating sub-module,

wherein the video processing sub-module is configured to acquire an input video frame corresponding to the input sound spectra, the input video frame comprising the multiple sound sources,

wherein the sound isolating sub-module is configured to isolate the predicted sound spectra from the input sound spectra by performing spectrum isolation processing on the input sound spectra according to the input video frame.

16. The device of claim 15,

wherein the video processing sub-module is configured to acquire a visual feature map according to the input video frame, the visual feature map comprising multiple visual feature vectors in k dimensions, each visual feature vector of the multiple visual feature vectors corresponding to one sound source in the input video frame,

wherein the sound isolating sub-module is configured to acquire k basic components according to the input sound spectra, the k basic components representing respective sound features in the input sound spectra, the k being a natural number; and acquire the predicted sound spectra according to a visual feature vector of the multiple visual feature vectors as well as the k basic components, a sound source of the predicted sound spectra being a sound source corresponding to the visual feature vector.

17. The device of claim 16, wherein the video processing sub-module is configured to implement:

outputting a video feature of the input video frame by inputting the input video frame to a feature extraction network; and

acquiring the visual feature map comprising the multiple visual feature vectors by performing max pooling on the video feature in time dimension.

18. The device of claim 16, wherein the sound isolating sub-module is configured to implement:

acquiring the predicted sound spectra as a dot product of a vector of the k basic components and the visual feature vector of k elements.

~~19. The device of claim 16, wherein the sound isolating sub-module is configured to implement:~~

~~acquiring a dot product of a vector of the k basic components and the visual feature vector of k elements;~~

~~acquiring a predicted mask by performing nonlinear activation processing on the dot product; and~~

~~acquiring the predicted sound spectra as a dot product of the predicted mask and initial input sound spectra for a first iteration.~~

~~20. The device of claim 16, wherein the sound isolating sub-module is configured to implement:~~

~~randomly selecting a visual feature vector from the multiple visual feature vectors; and~~

~~acquiring the predicted sound spectra according to the visual feature vector selected and the k basic components.~~

~~21. The device of claim 16, wherein the sound isolating sub-module is configured to implement:~~

~~selecting, from the multiple visual feature vectors, a visual feature vector corresponding to a loudest sound source; and~~

~~acquiring the predicted sound spectra according to the visual feature vector selected and the k basic components.~~

~~22. The device of claim 21, wherein the sound isolating sub-module is configured to implement:~~

~~acquiring a first dot product of a vector of the k basic components and the each visual feature vector of the multiple visual feature vectors;~~

~~acquiring a second dot product of the first dot product having been subject to nonlinear activation and initial input sound spectra for a first iteration;~~

~~acquiring average energy of the second dot product; and~~

~~selecting a visual feature vector corresponding to a location of max average energy.~~

~~23~~. The device of any one of claims 14 to 15~~22~~, further comprising a spectrum adjusting module configured to implement:

acquiring a residual mask according to the predicted sound spectra and historical cumulative spectra, the historical cumulative spectra being a sum of historical predicted sound spectra isolated during sound isolation;

acquiring residual spectra based on the residual mask and the historical cumulative spectra; and

acquiring complete predicted sound spectra as a sum of the residual spectra and the predicted sound spectra.

~~24. The device of claim 23, wherein the spectrum updating module is configured to implement:~~

~~acquiring the updated input sound spectra by removing the complete predicted sound spectra from the input sound spectra, wherein the sum of the historical predicted sound spectra comprises a sum of historical complete predicted sound spectra.~~

17~~25~~. The device of any one of claims 14 to 16~~24~~, wherein the spectrum isolating module is configured to implement,

in response to average energy of the updated input sound spectra being less than a preset threshold, determining that the input sound spectra comprise no sound spectrum.

18~~26~~. Electronic equipment, comprising memory and a processor, wherein the memory is configured to store computer instructions executable by the processor, wherein the processor is configured to implement the method of any one of claims 1 to 13 when executing the computer instructions.

19~~27~~. A computer-readable storage medium, having stored thereon a computer program which, when executed by a processor, implements the method of any one of claims 1 to 13.

20~~28~~. A computer program, when executed by a processor, implementing the method of any one of claims 1 to 13.