CLAIMS

1. A computer-implemented method for building at least one dataset for automatic speech recognition, the method comprising

generating an input dataset;

obtaining an audio dataset associated with the input dataset;

verifying the audio dataset using an audio verification module;

assigning at least one pseudo-label to the verified audio dataset;

storing the at least one pseudo-labelled audio dataset in a training database;

repeating the above steps until a pre-determined number of pseudo-labelled audio dataset have been stored in the training database; and

training an automatic speech recognition module based on the pre-determined number of pseudo-labelled audio dataset,

wherein the input dataset comprises a text dataset including a first text dataset and a second text dataset, resulting in a first output data and a second output data respectively, wherein the first text dataset and the second output data are configured to be input to a first inference module, and the second text dataset and the first output data are configured to be input to a second inference module.

2. The computer-implemented method of claim 1, wherein generating the text dataset comprises introducing at least one of a relatively uncommon vocabulary and a relatively uncommon language in the text dataset.

3. The computer-implemented method of claim 2, further comprises comparing the text dataset for speaker verification with an output transcription file of the automatic speech recognition module, and identifying at least one unsatisfactory result from the comparison.

4. The computer-implemented method of claim 3, wherein identifying the at least one unsatisfactory trained result includes determining a precision rate and/or a recall rate associated with one or more words in the text dataset.

5. The computer-implemented method of claim 3 or 4, further comprising selecting a speaker from a user group historically associated with accurate text-dependent speaker verification, and assigning the at least one pseudo-label to the audio dataset generated by the speaker without verification.

6. The computer-implemented method of any one of the preceding claims, further comprising generating an output transcription file by the automatic speech recognition module.

7. The computer-implemented method of claim 1, wherein the first output data and the second output data each comprises a transcription file.

8. The computer-implemented method of any one of the preceding claims, wherein the output of first inference module and the output of the second inference module are compared.

9. A device for building a dataset for automatic speech recognition, the device comprises

a processor configured to repeatedly:

generate an input dataset for audio verification;

obtain an audio dataset based on the generated input dataset;

verify the audio dataset using an audio verification module;

assign at least one pseudo-label to the verified audio dataset; and

store the at least one pseudo-labelled audio dataset in a training database until a pre-determined number of pseudo-labelled audio datasets have been stored in the training database,

wherein the input dataset comprises a text dataset including a first text dataset and a second text dataset, resulting in a first output data and a second output data respectively, wherein the first text dataset and the second output data are configured to be input to a first inference module, and the second text dataset and the first output data are configured to be input to a second inference module.

10. The device of claim 9, wherein the processor is configured to introduce at least one of a relatively uncommon vocabulary and a relatively uncommon language in the generation of the text dataset.

11. The device of claim 10, wherein the processor is configured to compare the text dataset for speaker verification with an output transcription file of the automatic speech recognition module, and identify at least one unsatisfactory result from the comparison.

12. The device of claim 11, wherein the processor is configured to identify the at least one unsatisfactory trained result based on a precision rate and/or a recall rate associated with one or more words in the output transcription file.

13. The device of claim 11 or 12, wherein the processor is configured to select a speaker from a user group historically associated with accurate speaker verification, and the processor is further configured to assign the at least one pseudo-label to the audio dataset generated by the speaker, without verification.

14. The device of claim 11, wherein the processor is configured to generate the text dataset based on feedback of the at least one unsatisfactory result.15. The device of claim 9, wherein the first output data and the second output data each comprises a transcription file.

16. The device of claims 9 to 15, wherein the output of first inference module and the output of the second inference module are compared.

17. A non-transitory computer-readable storage medium comprising instructions, which, when executed by one or more processors, cause the execution of the method for building a dataset for automatic speech recognition according to any one of claims 1 to 8.

18. A data processing device configured to perform the method of any one of claims 1 to 8.

19. A computer executable code comprising instructions building a dataset for automatic speech recognition according to any one of claims 1 to 8.

20. An automatic speech recognition module trained by the method of any one of claims 1 to 8.