

CLAIMS

1. A system comprising:
a processing unit; and
a storage device including program code that when executed by the processing unit causes the system to:
 - determine a first batch size for processing an audio signal;
 - collect a first batch comprising a first number of raw acoustic feature frames of the audio signal, the first number equal to the first batch size;
 - input the first batch to a speech recognition network;
 - determine a second batch size based on a word hypothesis output by the speech recognition network, the second batch size greater than the first batch size;
 - collect a second batch comprising a second number of acoustic feature frames of the audio signal, the second number equal to the second batch size; and
 - input the second batch to the speech recognition network.
2. A system according to Claim 1, wherein determination of the second batch size based on the word hypothesis comprises detection of a first non-silent word hypothesis output by the speech recognition network and determination of the second batch size in response to the detection.
3. A system according to Claim 1, the program code that when executed by the processing unit causes the system to:
 - collect a plurality of batches, each of the plurality of batches comprising the first number of acoustic feature frames of the audio signal;
 - input the plurality of batches to the speech recognition network until the word hypothesis is output by the speech recognition network.
4. A system according to Claim 1, the program code that when executed by the processing unit causes the system to:
 - detect an end-of-speech frame; and
 - reduce a size of a next collected batch of acoustic feature frames to include the end-of-speech frame as a last frame of the next collected batch.
5. A system according to Claim 1, wherein the first batch size comprises a first number of look-ahead frames and the first batch comprises first look-ahead frames, the number of first look-ahead frames of the first batch equal to the first number of look-ahead frames, the program code that when executed by the processing unit causes the system to:
 - collect a third batch comprising the first number of acoustic feature frames of the

audio signal, a time period associated with a plurality of frames of the third batch overlapping a time period associated with the first look-ahead frames of the first batch,

wherein determination of the second batch size comprises determination of a second number of look-ahead frames, and

wherein the second batch comprises second look-ahead frames, the number of second look-ahead frames of the second batch equal to the second number of look-ahead frames.

6. A system according to Claim 5, wherein the speech recognition network comprises a latency-controlled bi-directional long short-term memory model.

7. A system according to Claim 1, wherein the first batch size comprises a first number of look-ahead frames and the first batch comprises first look-ahead frames, the number of first look-ahead frames of the first batch equal to the first number of look-ahead frames, the program code that when executed by the processing unit causes the system to:

collect a third batch comprising the first number of acoustic feature frames of the audio signal, a time period associated with a plurality of frames of the third batch overlapping a time period associated with the first look-ahead frames of the first batch,

wherein the second batch comprises second look-ahead frames, the number of second look-ahead frames of the second batch equal to the first number of look-ahead frames.

8. A system according to Claim 7, wherein the speech recognition network comprises a contextual layer-trajectory long short-term memory model.

9. A computer-implemented method comprising:

collecting a first batch of acoustic feature frames of an audio signal, the number of acoustic feature frames of the first batch equal to a first batch size;

inputting the first batch to a speech recognition network;

in response to detection of a word hypothesis output by the speech recognition network, collecting a second batch of acoustic feature frames of the audio signal, the number of acoustic feature frames of the second batch equal to a second batch size, the second batch size greater than the first batch size; and

inputting the second batch to the speech recognition network.

10. A method according to Claim 9, wherein detection of the word hypothesis comprises detection of a first non-silent word hypothesis output by the speech recognition network.

11. A method according to Claim 9, further comprising:
detecting an end-of-speech frame;
in response to detecting the end-of-speech frame, collecting a third batch of acoustic feature frames of the audio signal, the third batch to include the end-of-speech frame as a last frame of the third batch; and
inputting the third batch to the speech recognition network.
12. A method according to Claim 9, wherein the first batch size comprises a first number of look-ahead frames and the first batch comprises first look-ahead frames, the number of first look-ahead frames of the first batch equal to the first number of look-ahead frames, the method further comprising:
collecting a third batch comprising the first number of acoustic feature frames of the audio signal, a time period associated with a plurality of frames of the third batch overlapping a time period associated with the first look-ahead frames of the first batch,
wherein the second batch comprises second look-ahead frames, the number of second look-ahead frames of the second batch equal to the second number of look-ahead frames.
13. A method according to Claim 12, wherein the speech recognition network comprises a latency-controlled bi-directional long short-term memory model.
14. A method according to Claim 9, wherein the first batch size comprises a first number of look-ahead frames and the first batch comprises first look-ahead frames, the number of first look-ahead frames of the first batch equal to the first number of look-ahead frames, further comprising:
collecting a third batch comprising the first number of acoustic feature frames of the audio signal, a time period associated with a plurality of frames of the third batch overlapping a time period associated with the first look-ahead frames of the first batch,
wherein the second batch comprises second look-ahead frames, the number of second look-ahead frames of the second batch equal to the first number of look-ahead frames.
15. A method according to Claim 14, wherein the speech recognition network comprises a contextual layer-trajectory long short-term memory model.