CLAIMS

1. A text-to-speech synthesis system, comprising:

a speech engine;

a processing unit; and

a neural network;

wherein, in a training mode:

the speech engine is configured to generate synthetic speech data for a first input text;

the processing unit is configured to compare the synthetic speech data to recorded reference speech data corresponding to the first input text, the processing unit further configured to extract at least one feature indicative of at least one difference between the synthetic speech data and the recorded reference speech data based on the comparison of the synthetic speech data to the recorded reference speech data; and

the neural network is configured to train based on, at least in part, the at least one feature extracted, the neural network further configured to generate a speech gap filling model based on, at least in part, the training, to fill in gap or differences between the synthetic speech data for the first input text and the recorded reference speech data corresponding to the first input text, and

wherein, in a synthesis mode:

the speech engine is further configured to generate speech output for a second input text based on, at least in part, the speech gap filling model;

the speech engine is further configured to generate an interim set of parameters for the second input text;

the processing unit is further configured to process the interim set of parameters based on, at least in part, the speech gap filling model to generate a final set of parameters; and

the speech engine is further configured to generate the speech output for the second input text based on, at least in part, the final set of parameters.

2. The text-to-speech synthesis system of claim 1, wherein the text-to-speech synthesis system is a parametric text-to-speech synthesis system.

3. The text-to-speech synthesis system of in claim 1, wherein the synthetic speech data, as generated by the speech engine, is based on, at least in part, at least one of a parametric acoustic model and a linguistic model pre-configured for a speaker.

4. The text-to-speech synthesis system of claim 3, wherein the synthetic speech data, as generated by the speech engine, is further based on, at least in part, the recorded reference speech data pre-recorded by the speaker.

5. The text-to-speech synthesis system of claim 1, wherein in the training mode, the processing unit is further configured to align the synthetic speech data and the recorded reference speech data preceding the comparison.

6. The text-to-speech synthesis system of claim 5, wherein the processing unit is further configured to implement one or more of pitch shifting, time normalization, and time alignment between the synthetic speech data and the recorded reference speech data.

7. The text-to-speech synthesis system of claim 1, wherein the at least one feature extracted include a sequence of excitation vectors corresponding to the at least one difference between the synthetic speech data and the recorded reference speech data for the first input text.

8. The text-to-speech synthesis system of claim 1, wherein in an update mode, the processing unit is further configured to:

compare the speech output for the second input text to a recorded reference speech data corresponding to the second input text; and

extract an updated at least one feature indicative of at least one difference between the speech output for the second input text and the recorded reference speech data corresponding to the second input text based on, at least in part, the comparison of the speech output for the second input text to the recorded reference speech data corresponding to the second input text.

9. The text-to-speech synthesis system of claim 8, wherein the neural network is further configured to update based on, at least in part, the updated at least one feature extracted, and the neural network is further configured to update the speech gap filling model based on, at least in part, the training.

10. A text-to-speech synthesis method, comprising:

in a training mode:

generating synthetic speech data for an input text;

comparing the synthetic speech data to recorded reference speech data corresponding to the input text;

extracting at least one feature indicative of at least one difference between the synthetic speech data and the recorded reference speech data based on, at least in part, the comparison of the synthetic speech data to the recorded reference speech data;

generating a speech gap filling model based on, at least in part, the at least one feature extracted to fill in gap or differences between the synthetic speech data for the input text and the recorded reference speech data corresponding to the input text; and

in a synthesis mode:

generating a speech output for a second input text based on, at least in part, the speech gap filling model, wherein generating the speech output comprises:

generating an interim set of parameters;

processing the interim set of parameters based on, at least in part, the speech gap filling model to generate a final set of parameters; and

generating the speech output for the second input text based on, at least in part, the final set of parameters.

11. The text-to-speech synthesis method of claim 10, wherein the synthetic speech data generated is based on, at least in part, at least one of a parametric acoustic model and a linguistic model pre-configured for a speaker.

12. The text-to-speech synthesis method of claim 10, wherein the synthetic speech data generated is further based on, at least in part, the recorded reference speech data pre-recorded by a speaker.

13. The text-to-speech synthesis method of claim 10 further comprising aligning the synthetic speech data and the recorded reference speech data preceding the comparison.

14. The text-to-speech synthesis method of claim 13, wherein aligning the synthetic speech data and the recorded reference speech data comprises implementing one or more of pitch shifting, time normalization, and time alignment between the synthetic speech data and the recorded reference speech data.

15. The text-to-speech synthesis method of claim 10 further comprising training a neural network based on, at least in part, the at least one feature to generate the speech gap filling model.

16. The text-to-speech synthesis method of claim 10 further comprising:

comparing the speech output generated for the second input text to recorded reference speech data corresponding to the second input text; and

extracting an updated at least one feature indicative of at least one difference between the speech output generated for the second input text and the recorded reference speech data corresponding to the second input text based on, at least in part, the comparison of the speech output for the second input text to the recorded reference speech data corresponding to the second input text.

17. The text-to-speech synthesis method of claim 16 further comprising updating the speech gap filling model based on, at least in part, the updated at least one feature.

18. A computer program product residing on a computer readable storage medium having a plurality of instructions stored thereon which, when executed across one or more processors, causes at least a portion of the one or more processors to perform operations comprising:

generating synthetic speech data for an input text;

comparing the synthetic speech data to recorded reference speech data corresponding to the input text;

extracting at least one feature indicative of at least one difference between the synthetic speech data and the recorded reference speech data based on, at least in part, the comparison of the synthetic speech data to the recorded reference speech data;

generating a speech gap filling model based on, at least in part, the at least one feature extracted; and

generating a speech output based on, at least in part, the speech gap filling model, wherein generating the speech output comprises:

generating an interim set of parameters;

processing the interim set of parameters based on, at least in part, the speech gap filling model to generate a final set of parameters; and

generating the speech output based on, at least in part, the final set of parameters.