CLAIMS

1. A computer-implemented method, comprising:

identifying, by a first computing device that is configured to respond to a particular, predefined hotword, a second computing device that is configured to respond to the particular, predefined hotword;

receiving, by the first computing device, audio data for an utterance spoken by a user;

determining, by the first computing device, that the utterance includes the particular, predefined hotword;

determining, by the first computing device, that the second computing device will output an initial response to the utterance;

receiving, by the first computing device, a transcription of the initial response to the utterance, wherein the initial response comprises synthesized speech audibly outputted by the second computing device;

based on the transcription of the initial response and based on the utterance, generating, by the first computing device, a subsequent response to the initial response; and

providing, for output by the first computing device, audio data for the subsequent response or a transcription of the subsequent response.

2. The method of claim 1, wherein providing, for output, audio data for the subsequent response or a transcription of the subsequent response comprises:

providing, to a speech synthesizer of the first computing device, an initial portion of the transcription of the subsequent response; and

providing, to the second computing device, (i) a remaining portion of the transcription of the subsequent response and (ii) instructions to audibly output synthesized speech of the remaining portion of the transcription of the subsequent response.

3. The method of claim 1, wherein:

determining that the utterance includes the particular, predefined hotword comprises:

determining that the utterance includes the particular, predefined hotword without preforming speech recognition on the audio data; and

determining that the second computing device will output an initial response to the utterance comprises:

receiving data indicating that the second computing device will output the initial response to the utterance.

4. The method of claim 1, wherein providing, for output, audio data for the subsequent response or a transcription of the subsequent response comprises:

providing the transcription of the subsequent response to a speech synthesizer.

5. The method of claim 4, wherein synthesized speech of a transcription of the subsequent response is received by a third computing device that is configured to generate a response based on the synthesized speech of the transcription of the subsequent response, the transcription of the initial response, and the utterance.

6. The method of claim 1, comprising:

after providing, for output, the transcription that corresponds to the response, receiving, by the first computing device, audio data of a second utterance;

based on the transcription of the initial response, based on the utterance, and based on the second utterance, generating, by the first computing device, an additional transcription of an additional response to the second utterance; and

providing, for output by the first computing device, the additional transcription.

7. The method of claim 1, wherein providing, for output, audio data for the subsequent response or a transcription of the subsequent response comprises:

providing the transcription to a display of the first computing device.

8. The method of claim 3, wherein receiving data indicating that the second computing device will output the initial response to the utterance comprises:

receiving, from the second computing device, a short range radio signal that indicates that the second computing device will output the initial response to the utterance;

receiving, from the second computing device and through a local network, data indicating that the second computing device will output the initial response to the utterance; or

receiving, from a server, data indicating that the second computing device will output the initial response to the utterance.

9. The method of claim 3, wherein determining that the utterance includes the particular, predefined hotword without preforming speech recognition on the audio data comprises:

extracting audio features of the audio data for the utterance;

generating a hotword confidence score by processing the audio features;

determining that the hotword confidence score satisfies a hotword confidence threshold; and

based on determining that the hotword confidence score satisfies a hotword confidence threshold, determining that the utterance includes the particular, predefined hotword.

10. The method of claim 1, wherein generating a subsequent response to the initial response comprises:

determining user information that is associated with a first user of the first computing device or with a second user of the second computing device; and

wherein generating the subsequent response to the initial response is based on the user information.

11. The method of claim 1, wherein generating a subsequent response to the initial response comprises:

accessing data that is associated with the initial response; and

wherein generating the subsequent response to the initial response is based on the accessed data.

12. The method of claim 1, comprising:

determining, by the first computing device, a location of the first computing device,

wherein generating the subsequent response to the initial response is based on the location of the first computing device.

13. The method of claim 3, comprising:

in response to receiving data indicating that the second computing device will output the initial response to the utterance, providing, by the first computing device and to the second computing device or to a server, the audio data of the utterance.

14. The method of claim 1, comprising:

generating, by the first computing device, a first audio fingerprint of the audio data for the utterance;

receiving, by the first computing device and from the second computing device, a second audio fingerprint of the audio data of the utterance;

comparing, by the first computing device, the first audio fingerprint to the second audio fingerprint; and

based on comparing the first audio fingerprint to the second audio fingerprint, determining, by the first computing device, that the audio data received by the first computing device represents the same utterance as the audio data received by the second computing device.

15. The method of claim 1, wherein the first computing device and the second computing device are configured to communicate with each other using short range radio.

16. The method of claim 1, wherein the first computing device and the second computing device are co-located.

17. The method of claim 1, wherein receiving a transcription of the initial response to the utterance comprises:

receiving, from the second computing device, the transcription of the initial response to the utterance.

18. A system comprising:

one or more computers and one or more storage devices storing instructions that are operable, when executed by the one or more computers, to cause the one or more computers to perform operations comprising:

identifying, by a first computing device that is configured to respond to a particular, predefined hotword, a second computing device that is configured to respond to the particular, predefined hotword;

receiving, by the first computing device, audio data for an utterance spoken by a user;

determining, by the first computing device, that the utterance includes the particular, predefined hotword;

determining, by the first computing device, that the second computing device will output an initial response to the utterance;

receiving, by the first computing device, a transcription of the initial response to the utterance, wherein the initial response comprises synthesized speech audibly outputted by the second computing device;

based on the transcription of the initial response and based on the utterance, generating, by the first computing device, a subsequent response to the initial response; and

providing, for output by the first computing device, audio data for the subsequent response or a transcription of the subsequent response.

19. A non-transitory computer-readable medium storing software comprising instructions executable by one or more computers which, upon such execution, cause the one or more computers to perform operations comprising:

identifying, by a first computing device that is configured to respond to a particular, predefined hotword, a second computing device that is configured to respond to the particular, predefined hotword;

receiving, by the first computing device, audio data for an utterance spoken by a user;

determining, by the first computing device, that the utterance includes the particular, predefined hotword;

determining, by the first computing device, that the second computing device will output an initial response to the utterance;

receiving, by the first computing device, a transcription of the initial response to the utterance, wherein the initial response comprises synthesized speech audibly outputted by the second computing device;

based on the transcription of the initial response and based on the utterance, generating, by the first computing device, a subsequent response to the initial response; and

providing, for output by the first computing device, audio data for the subsequent response or a transcription of the subsequent response.