Takeaway Questions for Automatic Speech Recognition (ASR)

Broad questions

1. What is the input and what is the output? How is the input speech signal represented? How is the output represented? Give a statistical formulation of ASR system in terms of how the input and output are represented.
2. Given "only" two instances of speech utterances, how can we decide automatically whether those two instances correspond to the same word/phrase or not?
3. How can each word in the ASR system be modelled? What kind of prior knowledge is needed?
4. What kind of model is needed to model the relationship between the acoustic speech signal/feature representation and the linguistic units, such as, phones?
5. How do we match an acoustic speech signal with a word hypothesis?
6. How can we model the set of possible word hypotheses in an ASR system? What kind of prior knowledge is needed? Hint: how can the relationship between the words be modeled? What kind of model?
7. What is the difference between isolated word recognition task, connected word recognition task and continuous speech recognition task?
8. What is the role of decoder in an ASR system? What does it involve?

Subword unit-based ASR system

1. Why subword unit-based word modeling is preferred over whole word-based modeling?
2. What are the different subword units that are suitable for building automatic speech recognition system? What are phonemes/phones and graphemes? What kind of prior knowledge is needed when using phones or graphemes to model words in a speech recognition system? How is the prior knowledge integrated into the speech recognition system?
3. What is context-independent subword unit modeling? What is context-dependent subword unit modeling? Why is context-dependent subword unit modeling preferred over context-independent subword unit modeling? What are the challenges in context-dependent subword unit modeling?
4. What is pronunciation variation? How can we handle the pronunciation variation problem?

Instance-based ASR

1. What are the advantages and disadvantages of instance-based ASR?
2. In what kind of scenarios, instance-based ASR could be put to practice?
3. In instance-based ASR, how do you choose local cost and local constraints for to match two instances using dynamic programing?

Language modeling

1. What is the linguistic unit (e.g., phone, word, phrase) used for language modeling?
2. What is the goal of language modeling in speech recognition system? What are the different language modeling methods?
3. Do all words in a language have same prior probability? Would communication be efficient if all words were equiprobable and can follow each other in an equiprobable manner? What kind of Markov model would reflect that scenario?
4. What kind of resources are needed to develop a language model?
5. In n-gram modeling, how can low count words/contexts be effectively modeled? How can unseen contexts be handled?