Text-to-speech Recap Questions

1. Difference between automatic speech recognition and text-to-speech synthesis
    a. Input and output
    b. Challenges
    c. Theoretical criteria
    d. Evaluation
2. Natural language processing in text-to-speech synthesis
    a. Goal
    b. How do we handle abbreviations and numbers?
    c. How do we get phonetic representations?
    d. How can we get intonation and duration information?
3. Concatenative text-to-speech synthesis
    a. Basic idea
    b. What is the distinction between diphone synthesis approach and unit selection speech synthesis approach?
    c. What is the cost function for unit selection speech synthesis and how is it optimized?
4. HMM-based (Statistical parametric) speech synthesis
    a. Basic idea
    b. How are the intonation and duration information estimated/modeled?
    c. How is the vocal tract system information estimated?
    d. What is the purpose of vocoding?
    e. What are the key differences between HMM-based speech synthesis and unit selection speech synthesis? How those two approaches can be combined?
5. Neural text-to-speech synthesis
    a. In HMM-based speech synthesis, where can we apply neural networks?
    b. Where does Wavenet mainly differs in comparison to HMM-based speech synthesis and how?
    c. Describe briefly the "basic" idea of Tacotron neural-based TTS approach. What is the main difference between Tacotron and other TTS approaches such as, concatenative synthesis, HMM-based speech synthesis?
6. In the speech recognition part, we observed that the problem of matching an acoustic signal with a word hypothesis can be formulated as four sub-problems. How can we formulate concatenative synthesis, HMM-based synthesis and Wavenet approach along those lines? What additional sub-problem you may need for TTS?