DT2112 home exam spring 2019

Assignments

1. Automatic speech recognition and human perception

1.1. E-level assignment

Describe briefly what the following ASR resources/processes are and how they are used in standard HMM-based ASR:

- the parametrization (1p),
- the acoustic models (1p),
- and the language models (1p).

Also describe what an N-best list is, and why developers may prefer to get an N-best list as an ASR result instead of a single best guess (1p).

1.2 C-level assignment

Discuss the ASR resources/processes in the E-level assignment (the parametrization, the acoustic models, the language models, and the N-best list; 1p each) in terms of humans: how do they relate to human perception?

1.3. A-level assignment: Semantic context and perception

Describe two different ways in which something that is not directly a part of a speech sound (a phoneme or a word) can affect our perception of it, and how this effect takes place (2p). You may suggest anything, including eternal sounds, visual events, events that took place long before, just before, or after the speech sound in question was spoken... For each of your two suggestions, discuss briefly how it would affect automatic speech recognition, in comparison with human perception (2p).

2. Synthesis and production

2.1. E-level assignment

Describe briefly the following four types of speech synthesis, with a special focus on how they differ: formant synthesis (1p), diphone synthesis (1p), unit selection synthesis (1p) and HMM based synthesis (HTS; 1p).

2.2. C-level assignment

Discuss how the following phenomena, which are all frequently occurring in human speech, relate to unit selection synthesis: co-articulation (1p), reduction (1p), hesitation - for example a pause or an "ehm" (1p), and correction - for example the repetition of a word that "came out wrong" (1p). For each phenomenon, discuss if it finds its way into unit selection synthesis at all, and if so what the effects of that might be, as compared to the effects it has in human speech.

2.3. A-level assignment

This is a theoretical assignment, in which you give suggestions of how the pause and resumption of speech can be simulated in unit selection synthesis. The context is as follows:

A university level student is listening to a text book with synthetic speech. The speech synthesis can be voice controlled, that is you can go to the next section, adjust speed, pause, stop etc. with voice commands. The phone rings and the students says "Stop!" to stop the synthesis.

How should the synthesis handle the sudden break?

Discuss a few alternatives of how and at which point in its reading the system should stop (1p), and how and from which point the synthesis should then resume speaking when the student tells it to continue (1p).

Give 3-5 suggestions of what you can do to simulate a human-like pause and resumption in a unit selection synthesis. Explain why you think each of your suggestions is a good idea: in what way are they human-like? Suggest shortly how you think the methods might be implemented. (2p)

3. Dialogue

3.1. E-level assignment

Describe and discuss briefly at least 4 ways in which a current standard spoken dialogue system performing a well-defined dialogue task differs from a human performing the same task. Exemplify by describing a plausible system behaviour and its human counterpart. (4p)

3.2. C-level assignment

One way on viewing a spoken dialogue system is as if it was a straightforward replacement for some other user interface, such as a web form or a DTMF tone control. Another view is that it is a substitute for a human helper such as a travel agent. Suggest and motivate, using existing applications and tasks as examples, 1 application where the "interface replacement" view makes more sense (1p), 1 application where the "human substitute" view makes more sense (1p). Also suggest and discuss the pros and cons for 2 applications where the choice would be less obvious (2p).

3.3. A-level assignment

The objective of this assignment is for you to interact with an existing state-of-the-art spoken dialogue system and judge its components, performance, user friendliness etc.

Talk to a commercial spoken information system, test it and make a report on its strengths and weaknesses. You can choose any Swedish or non-Swedish spoken dialogue system you want.

To get a fair idea of the system's performance you should call several times and try different types of user input, by varying what is said to the system (e.g. stations or times that are easy or difficult to recognize, in the case of an information service) and how it is said (clearly, normally, casually). You may want to record your calls, in order to analyse them later on.

Report the following, briefly:

Describe the system. What components does it have, and how do they influence the dialogue? Also give a brief description of the tasks you tested. (1p)

Describe the interaction between the user and the system: How free is the dialogue; can the user choose what to say when? Can the user interrupt the system? How does the system confirm what the user has said? (1p)

Describe the error management. What happens when something goes wrong? Is it easy or difficult to correct misunderstandings? How far back in the dialogue do you have to return to repair? (1p)

Discuss the user friendliness the system: Is the system easy to use: is it e.g. evident from the start what you should do to get the information? Is the system efficient: how long does it take to get the information? Do you like using it? (1p)

(Subtraction of 1 p for irrelevant or erroneous answers.)

4. Data collection

4.1. E-level assignment

Describe the steps you would go through to make a data collection (½ p per step up to 4p. Subtraction of ½ p for irrelevant or erroneous steps).

4.2. C-level assignment

Find the descriptions of the following corpora on the web: Switchboard, Columbia Games corpus and D64. Answer the following questions for each corpus:

How long are they, how many different speakers do they contain, are speakers repeated, and approximately how much time per speaker? What languages are represented? (1p)

Characterize the type of speech they contain. Read, scripted, spontaneous? Monologue or dialogue (and number of participants)? What is the task given to the speakers, if any? (1p)

Describe the purpose of the corpus. What is it intended to be used for? If it is a general purpose corpus, discuss briefly what you think it would be most suitable for. (1p)

Pick one characteristic from each corpus that you think is most different from the other of the five corpora - a characteristic that makes it stick out. Suggest briefly if and when that characteristic may be important for those using the corpus. (1p)

(Subtraction of 1 p for irrelevant or erroneous answers.)

4.3. A-level assignment

Corpora are recorded for a wide range of reasons, from the very specific purpose-build corpora to broad, general exploratory corpora. The design decisions that go into a corpus collection are largely dependent on the purpose of the corpus. At the end of this question you'll find a list of characteristics and design considerations for corpus collection. Select the three most important features from the list, and motivate your selection briefly, for each of:

- A corpus that is intended for training of acoustic models for speech recognition of street names in an in-car environment (1p)
- A corpus that is intended for broad vocabulary unit selection synthesis of academic literature (1p)
- A corpus that is intended to provide insights into the semantic concepts involved a spoken dialogue system that guides visitors to a museum (1p)
- A corpus that is intended to allow scientific studies of breath, posture, and speech for turn-taking in human face-to-face interaction (1p)

Choose from:

- Audio quality
- Video quality
- Synchronization of different recordings (e.g. audio channels; audio and video)
- Channel separation (e.g. one speaker per audio channel; one person per video track)
- Control over the subjects (e.g. the task performed)
- Control over the environment (e.g. noise, interruptions)
- An ecologically valid environment (i.e. recording takes place in an environment where the task would ormally take place)

- An ecologically valid task (i.e. the task is one that the subjects may perform under normal circumstances)
- Demographics (e.g. subjects selected so that they are representative of some relevant demographic group)
- Gender balance
- Age balance
- Control over the linguistic background of subjects (e.g. exclude second laguage speakers, certain dialects)
- Control over the subjects pre-recording behaviour (e.g. no alcohol or tobacco the previous day)
- Repeated recordings of the same subject (e.g. on different days or in different places)
- Recordings of a large number of subjects
- Extensive recordings of each subject
- Mobility of the subjects (e.g. make them sit quite still, or allow them to move around)
- The inclusion of (and control of) props (i.e. stuff in the recording environment, e.g. conversation pieces such as a teddy bear or a painting)
- Having a researcher present with the subjects during the recordings
- Monitoring the recordings
- Recording more than one person at the same time
- Controlled audio quality (specific microphones

(Subtraction of 1 p for irrelevant or erroneous answers.)

5. Evaluation

5.1. E-level assignment

Describe briefly the process of evaluation (in general or a specific type of system) by listing important steps in order (1 p per step up to 4p. Subtraction of 1 p for irrelevant or erroneous steps).

5.2. C-level assignment

Evaluations are made from the perspective of some party with an interest in the technology, the application or the service in question. The metric for the evaluation, that is what constitutes "good", is dependent on the requirements, desires and goals that this party has.

Describe two clearly different speech technology applications. For each of them, discuss what "good" would mean – what metric would be used and what data would be taken into consideration – at each of the following: a research institution (2*½p), a R&D company selling the underlying technology to service providers (2*½p), a company selling the service to end users (2*½p), and an end user (2*½p).

(Subtraction of ½-1 p for irrelevant or erroneous answers.)

5.3. A-level assignment

Do a literature search and find examples of 3 clearly different academic research evaluations relevant to speech technology. They should differ considerably in at least one of method or target/objective. Refer to the publications when you answer each of the following questions:

What is the main motivation behind this evaluation (3 * 1/3 p)? Describe the method – how is the evaluation performed (3 * 1/3p)? Does the evaluation involve human judges, automatic measures, or both (3 * 1/3p)? Critique – suggest something that could improve the evaluation (3 * 1/3p)?