Yohan
Pellerin

Home Exam : Speech technology

Question E level :

1)

- « Additionally, the signal is normalized by stretching or compression to compensate for variability in the durations of speech segments. » This is an erroneous statements, Normalization is the process of adjusting the amplitude or volume of the audio signal to a standard level. This can be important to ensure that the speech recognition system can handle variations in loudness and reduce the impact of differences in microphone sensitivity or distance from the speaker.

- « The final output of the system is often a logical description of actions and objects » this is an erroneous statements, the decoded output is the transcribed text that is generated by the speech recognition system

- « Using statistical models to match the acoustic features to phonetic units: Once the relevant acoustic features have been extracted, statistical models such as Hidden Markov Models (HMMs) or Neural Networks can be used to match these features to symbol sequences of phonemes. The models are trained using large datasets of speech and corresponding transcriptions to learn the relationships between the acoustic features. » This statement is one of the most correct statement, it is a very important step in speech recognition

- « Once the speech signal has been cleaned up, it is analyzed to extract relevant acoustic features. »   This statement is one of the most correct statement, feature extraction is a critical step in the process of speech recognition, which involves transforming the raw audio signal into a set of features that can be easily analyzed and processed by the computer.

2)

- « A third important aspect of speech production is respiration, which provides the air flow necessary for the production of speech. » This statement is incorrect, for example, non-pulmonic mechanisms involve alternative ways of generating airflow for speech production, often through manipulation of the vocal tract or initiation of airflow that doesn't strictly rely on the lungs.

- « Prosody is meaningful in tonal languages such as Mandarin Chinese, but conveys only attitude in other languages » This statement is incorrect, in non-tonal languages, prosody conveys not only attitude but also contributes to emotional expression, pragmatic meaning, speech boundaries and syntax.

- « Articulation:Oneimportantaspectofhumanspeechproductionisthe process of articulation, which involves the precise movements of the lips, tongue, and other articulators to produce speech sounds. » And « Phonation:Anotherimportantaspectofspeechproductionisphonation, which refers to the vibration of the vocal cords to produce sound. » are two important aspects of human speech production which make this two part the most important correct statements.

3)

- « The text output from the speech recognition engine is processed by a large language model, which maps the user's intent to the real world and identifies the relevant actions or information », the statement incorrectly suggests that the mapping of the user's intent to the real world and identifying relevant actions or information is solely performed by the language model. In reality, this process converts the sequence of words into a semantic representation that can be used by the dialogue manager.
- « Naturallanguagegeneration:Iftheresponseinvolvesgeneratingspeech,a natural language generation (NLG) module converts the text response into spoken words. », it does converts into spoken words but only words the spoken part is handle by the speech synthesis part.
- « The spoken input is processed by a speech recognition engine, which transcribes the speech into text. » this is a very important correct statement to be able to respond to the demand well.
- « Action management: The action plan is realised by an actuation module which controls real-world effects (e.g. through robotic actuators or home controls). » this is also an important correct statement, It takes a semantic representation of the user's text, figures out how text fits in the overall context and creates a semantic representation of the system response. This is determine the accuracy of the response

4)

- **Erroneous Statement 1** "In most cases, using an already existing dataset is going to be sufficient to answer complex research questions". This statement is not universally true. While existing datasets can be valuable resources, they may not always be sufficient or appropriate for addressing complex research questions, especially if the research requires specific data that is not available in existing datasets. Additionally, the suitability of existing datasets depends on factors such as data quality, relevance, and alignment with the research objectives.
- **Erroneous Statement 2,** "This involves pre-testing the data collection tools and techniques on the participants...". Piloting the data collection method involves testing the procedures and instruments, not the participants themselves. Piloting ensures that the data collection process is feasible, efficient, and yields the desired information without placing undue burden on participants. Testing the data collection tools and techniques helps identify any issues or challenges that need to be addressed before conducting the actual data collection on a larger scale.
- The two other part : « define the research question » and « collect the data » are correct statement and important one.

5)

- "In many cases, this can be controlled for by eliminating the human entirely and performing an evaluation without humans." This statement is erroneous, while it may be possible to automate certain aspects of speech technology evaluation and perform evaluations without human involvement to some extent, completely eliminating the human factor is often impractical or undesirable.

- "It's important to clearly define the metrics that will be used to evaluate the performance of the speech technology, such as accuracy, precision, and recall." This statement is erroneous, while accuracy, precision, and recall are indeed common evaluation metrics used in various contexts, they may not always be suitable or sufficient for evaluating speech technology.
- The two other part : « Define the goal of the evaluation » and « select the evaluation data » are correct statement and important one.

Question level C

1)

- "This step is not typically included in ASR." This statement incorrectly implies that top-down processing is not typically included in automatic speech recognition (ASR) systems. However, many modern ASR systems incorporate elements of top-down processing by utilizing language models, context, and prior knowledge to aid in speech recognition.
- "In both ASR and human speech perception, the next step is to recognize the patterns in the extracted features, such as identifying phonemes or words." the statement oversimplifies the process by suggesting that recognizing patterns directly leads to identifying phonemes or words. In reality, recognizing patterns involves more complex computational processes, such as statistical modeling, machine learning algorithms, or neural network architectures, which aim to map the extracted features to linguistic units (e.g., phonemes, words).
- " Auditory perception: The first step in both ASR and human speech perception is auditory perception, which involves processing the sound waves that make up speech." This is good statement, auditory perception is the first step in both automatic speech recognition (ASR) and human speech perception. In ASR, the raw audio signal undergoes preprocessing to enhance its quality and prepare it for further analysis. And for human, the brain analyzes various acoustic features of speech, such as pitch, duration, and intensity, to extract linguistic information.
- "Segmentation: In both ASR and human speech perception, the next step is to segment the speech into meaningful units, such as words or phonemes." Humans naturally segment continuous speech into meaningful units, such as words, syllables, and phonemes. Segmentation in automatic speech recognition (ASR) refers to the process of dividing a continuous stream of speech into smaller, meaningful units such as words, phonemes, or other linguistic elements.

2)

- "Synthesized speech can usually convey the full range of emotions that humans can express through speech, but fails to match the emotional expression to the situation." This statement is incorrect because synthesized speech may not always be capable of conveying the full range of emotions that humans can express naturally. While advancements have been made in emotional speech synthesis, achieving human-like emotional expressiveness remains a significant challenge. Therefore, the assertion that synthesized speech can "usually convey the full range of emotions" is overly optimistic and not universally true.
- "Emotion expression: An other phenomen on that behaves differently in synthesized speech than in human speech is the expression of emotions. Synthesized speech can usually convey the full range of emotions that humans can express through speech, but fails to match the emotional expression to the situation. " The statement contains an erroneous assertion.

While it's true that synthesized speech technology has advanced significantly and can convey a range of emotions, the claim that synthesized speech can "usually convey the full range of emotions that humans can express through speech" is overly optimistic and not universally true.

- "Synthesized speech may struggle to accurately reproduce coarticulation, resulting in a less natural and sometimes unclear sound. This statement is correct, coarticulation is difficult for synthesized speech
- "One phenomenon that behaves differently in synthesized speech than it does in human speech is prosody, which refers to the rhythm, melody, and intonation of speech. Neural synthesized speech often lacks prosody, resulting in a monotonous or robotic sound." This statement is correct, in synthesized speech, prosody, which includes variations in pitch, duration, and intensity, may sound unnatural or robotic compared to human speech. Synthesized speech often lacks the subtle variations and nuances in intonation and rhythm that are characteristic of natural human speech.

3)

- "The channel capacity of Shannon's theory can also be applied to human conversation by examining the maximum amount of information that can be transmitted in a given period of time, such as during a conversation or over a phone call." This statement is correct, human conversation involves the transmission of information through spoken language, and like any communication channel, it has limitations in terms of the amount of information that can be effectively conveyed within a given timeframe.
- « The concept of entropy in Shannon's theory can be applied to the study of conversational turn-taking, as it can help to explain how speakers manage to take turns in a conversation without interrupting each other. » This statement is correct, entropy, in this context, refers to the unpredictability or uncertainty of a communication system. In conversational turn-taking, speakers must manage the uncertainty of when to start and stop speaking to avoid interruptions or overlaps.
- « Shannon's theory of information can be used to explain how the situation and the environment of a conversation are an integral part of the conversation. » This statement is incorrect, Shannon's theory primarily focuses on the transmission of information over communication channels, and while it provides insights into factors like noise and channel capacity, it doesn't directly address the role of situational context and environment in shaping conversation dynamics.
- « Shannon's theory can be used to explain how humans achieve common ground as an emergent result of the communal and simultaneous effort of conversation. ».This statement is incorrect, it's stretching the application of Shannon's theory to claim that it can explain how humans achieve common ground in conversation.

4)

- « In particular, the overall success of the meeting can only be judged by the meeting participants themselves. » This sentence is erroneous, Participants' perceptions can be influenced by individual preferences and biases, and survey responses may not always align with objective measures of meeting effectiveness.
- « Analysis of Meeting Content: In addition to analyzing the multimodal data, it is important to analyze the content of the meeting itself. An efficient method is to have the meeting participants taking notes on meeting topics, agenda items, and decisions made during the meeting. » This statement is incorrect. While understanding the content of the meeting, such

as topics discussed and decisions made, is crucial for assessing meeting effectiveness, it's equally important to consider participants' recollection of the meeting. In certain cases, what participants remember about the meeting holds greater significance than the meeting's actual content.

- « Collection of Multimodal Data: To gain insight into what makes for efficient distant meetings with video, it is important to collect multimodal data that captures both verbal and nonverbal communication. This can include video recordings of the meeting, as well as data on body language and facial expressions. » This statement is correct, collecting multimodal data can offer rich insights into communication dynamics, such as nonverbal cues and interaction patterns, this informations can be correlated to the efficiency of a meeting.
- « Selection of Participants: When collecting data for the purpose of learning what makes for efficient distant meetings with video, it is important to carefully select participants who have relevant experience with this type of meeting. This may involve selecting individuals who have participated in distant meetings before or who have experience with video conferencing. » This statement is correct, it wouldn't make sence to ask questions to people who never had experience the problem.

5)

- In « A corpus that is intended for training of acoustic models for speech recognition of street names in an in-car environment: » , « Control over the environment: Since the corpus is intended for in-carspeech recognition, controlling the environmental factors such as noise and interruptions is important to ensure the speech recordings are representative of the in-car environment. » The statements is incorrect, ensure the speech recordings are representative of the in-car environment correspod more to an ecologically valid environment
- In « A corpus that is intended to provide insights into the semantic concepts involved in a spoken dialogue system that guides visitors to a museum:

  An ecologically valid environment: Since the corpus is intended to provide insights into the semantic concepts involved in a spoken dialogue system guiding visitors in a museum, recording in an ecologically valid environment can ensure the speech recordings are representative of the real-world scenario. ». It is not one most important features, you don't need the audio to look like, it was a true guide but you need it to be clear without noise behind or interuption. Then, I would say that the Control over the environment (e.g. noise, interruptions) is much more important.
- For the corpus that is intended for broad vocabulary unit selection synthesis of academic literature, Extensive recordings of each subject, Control over the linguistic background of subjects and Control over the subjects pre-recording behavior are important features
- For corpus that is intended to allow scientific studies of breath, posture, and speech for turn-taking in human face-to-face interaction: Video quality, Synchronization of different recordings and Mobility of the subjects are important features