Example questions from the lecture with exemplary answers. No guarantee for correct answers.

1 Multi Modality

1. What does the term “modality” describe?

Solution: Multi Modal system:
Modality describes the communication channel that is used to acquire/convey information. It also covers the way an idea is expressed or perceived or the manner an action is performed.

Communication act:
Defines the type of data that is exchanged

2. Name the three main modalities and describe them briefly.

Solution:

- Visual: used in seeing, compare against optical
- Auditive: Related to the sense of hearing
- Tactile: Experienced by the sense of touch
- *Haptic*: Depending on the definition, haptic is a part of the tactile modality

3. What is a mode?

Solution: Multimodal system:
State that determines the context in which data is interpreted

Communication act:
Mode determines the context in which the data is interpreted

4. What is multi modality?

Solution: Communication or interaction through different modalities (channels) at the same time.

5. What is a multimedia system?

Solution: Multi media system: Focus on the medium/technology rather than the application or the user.

6. Why is multimodality important?

Solution: Noise
Multimodal interaction is part of the natural human behaviour
It increases

- Efficiency
- Redundancy
- Consistency
- Robustness

7. Difference between multimedia and multimodal system?

Solution: Both use multiple communication channels.
A multimodal system is able to automatically model the content of the information at a high level of abstraction.

8. What is multimodal interaction?

Solution: In a multi-modal system the multi-modal interaction is the communication with at least two modalities. The modes describe the communication channel.
Example: Audio Visual - Automatic Speech Recognition (AV-ASR)

9. Name 2 modalities of speech recognition (except speech) and explain them shortly.

Solution:

- Visual: e.g. lipreading, gestures
- Electromyography: tracking muscle movements with electrodes

10. What is a gesture?

Solution: Gestures are mostly hand movements that are aligned to speech.

(a) Are gestures part of speech?/ Why are gestures part of speech?

Solution:

- Gestures are part of speech. It has been shown that even blind people use it even in communication with other blind people.
- The more speech is absent the amount of gestures increases. In fully absent speech, gestures represent the sign language.
- Types of gestures

- Deictic gestures
  Create meaning by spatial reference
- Beat gestures
  Baton like movements, add emphasis to speech items that are important
- Representational gestures
  * Iconic: Expressing concrete physical properties
  * Metaphoric: Expressing a physical metaphor

11. Give an example for a multimodal system.

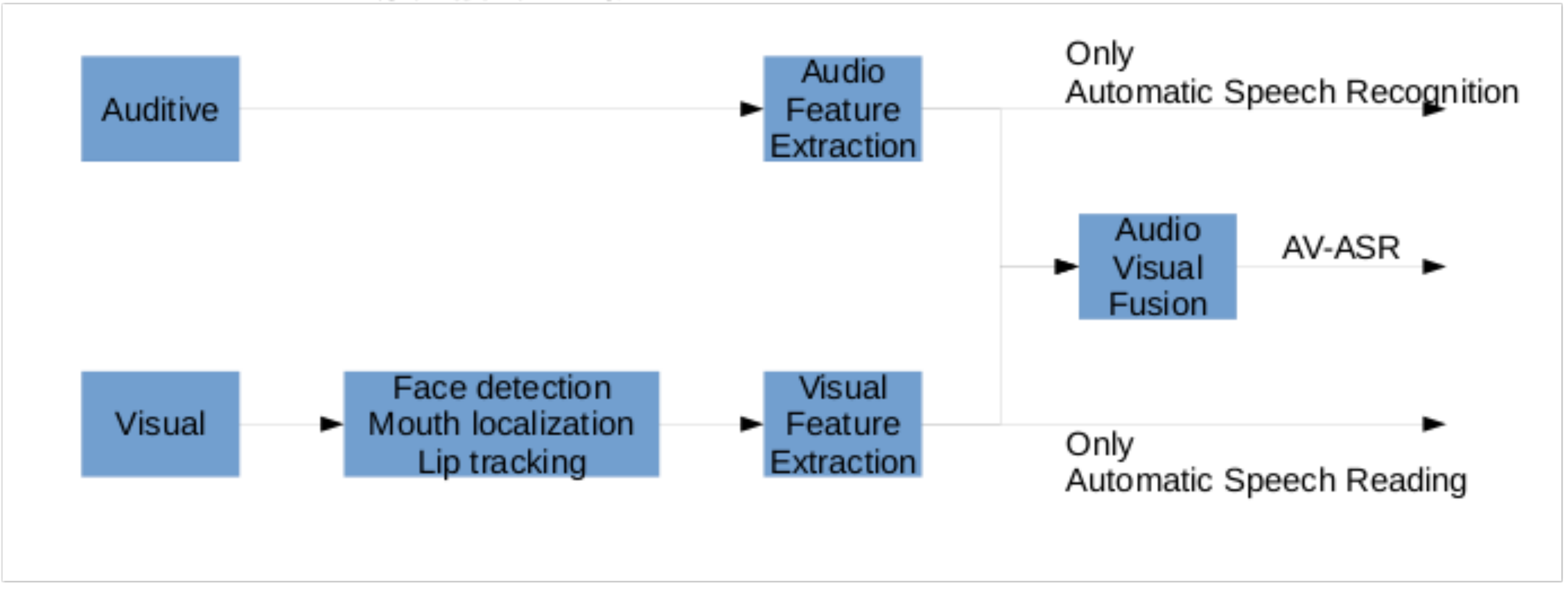
Solution: AV-ASR, Navigation System

12. What is AV-ASR?

Solution: Audio Visual Automatic Speech Recognition
Multi modal system for speech recognition

13. Draw the typical architecture for audio-visual speech recognition.

Solution: a



(a) Describe the components of AV-ASR

Solution:

- Auditive Input
  Audio Feature Extraction
- Visual Input
  Face detection
  Mouth localization
  Lip tracking

  Visual feature extraction

Both streams can be used separately:
Auditive: Only Automatic Speech recognition
Visual: Only Automatic Speech reading

Both streams combined: Audio Visual Fusion: AV-ASR

14. Why do we need AV-ASR?

Solution: Noise
Improve robustness compared to system with only one mode

15. Describe the constraints of AV-ASR

Solution:

- Auditive input
    - Background noise: System, Room, ...
    - Microphone distance: Echoes, phasing
    - Crosstalk: Hard to detect more than one speaker
- Visual input
    - Noise: Darkness, Color, ...
    - Movement: Speaker movement
- Feature wise
    - Speech style: Text read from paper, free talk (chatter), text commands, ...
    - Language Model and Acoustic model
    - Overfitting to training data

16. Explain visemes.

Solution: Basic classification unit for visual speech
Describes the mouth/face movement that is part of lip-reading.

17. Explain phonemes.

Solution: Basic classification unit for speech
There are 42 phonemes in American English.
Phonemes are generated by specific positions or movements in the vocal tract.

18. What are visual features?

Solution:

- Appearance based
  - Pixels contain information → Easier, more robust information
  - High dimensional input
- Model based features
  - Predefined model based parameters are used → low dimensionality
  - Model is more difficult to obtain

19. Name challenges of the visual part.

Solution:

- Feature selection
  Robust & invariant to lighting, motion, rotation
- Motion estimation
  Finding correspondence in a sequence of images
- System implementation
  Real time requirements aka needs to be fast
- *Additional: Active vision*
  Control camera (pan, tilt, zoom, ...)

20. Multi-modal fusion →HINT: Difference & Definition of early and late fusion is often answered wrong

(a) Name fusion strategies

Solution: Fusion strategies

- Early fusion (Feature level)
  - Integrate signals at feature level
  - Based on Hidden Markov Model (HMM)
  - used for closely coupled and in synchrony modalities, eg. speech & lip movement
  - Problem: Modes can differ substantially in information content or the time scale characteristics
    → Systems tend not to generalize well
  - Requires large amount of training data
  - Easy to implement
- Late fusion (Decision level)
  - Integrate information at semantic level
  - use individual recognizer, trained with uni-modal data
  - Easier to scale up

- Requires fine grained time stamping
  → Less data needed
- Harder to implement

(b) Where can it happen?

Solution:

- Feature level
  Very simple to be multimodal → does not work that well
- Phoneme/Viseme level
  More complicated
- Decision level
  Word level

(c) How to find weights?

Solution:

- Discriminatively trained weights
  Pre trained for fixed users and fixed environments
  → Not very dynamic
- Max entropy weights
  Measurement of change/ rate of change
  → Not every change results in useful information, eg. increasing overall loudness in a room results in a high change but does not contain more information
- SNR

21. Explain Electro-myography

Solution: Electromyography is another modality
In EM the muscle movement is measured with electrodes. For speech recognition these are the muscles that are part of the vocal tract. Only the external muscles are accessible.

(a) Reasons for electro-myography

Solution:

- Noisy environments: Dark, loud (eg. plane cockpits)
- Privacy concerns

(b) Problems with electro-myography

Solution:

- Frictional sounds can not be recognized
- Patterns vary with the arrangement of words

2 Machine Learning

1. What is machine learning?

Solution:

- Enable machines to act without explicitely programming it
- Predicts unkown from uncertain data
- Self configuring data structures
  →Allow a computer to do things that would be called intelligent if a human did it

2. What do we do with it? / What is it good for?

Solution:

- Prediction about (discovered) phenomenons in the data
- Description/ understanding data in a new way

3. How to implement it compared to classical software engineering?

Solution: Machine Learning makes decisions based on data instead of guessing.

Classical software engineering process

- Analyze/Specify: Interview experts and users to determine actions
- Design/Model: Apply computer science knowledge to design
- Implement: a solution
- Test

This fails when

- Requirements are hard to collect
- System must resolve difficult trade-offs

Examples

- No experts available
  → eg. telephone fraud

- Inarticualte experts
  → eg. handwriting recognition
- Requirements change very fast
  → eg. computer intrusion detection
- User dependent requirements
  → eg. E-Mail filtering

4. Explain the difference between machine learning and AI.

Solution: Machine learning is a part of AI

5. Explain the principle of feature extraction

Solution: Machines are limited in computational power & memory & storage.
→ Present stimulies using their distinctive attributes
Example apples & oranges features:

- shape
- size
- color

(a) Name and explain different forms of learning

Solution:

- Supervised learning
  Learn statistics that map input to output from examples which are labeled into known classes
- Unsupervised learing
  Identify patterns in examples but no label is provided. E.g. clustering

6. How is speech recognition performed in a machine learning context?

Solution: It is performed in the frequency domain. The distinctive feature is the frequency distribution for certain letters or words

7. Describe the Perceptrons principle

Solution:

- Simulates the brain cognitive mechanism
- Dense interconnection of computing cells based on the human brain
- Inputs are multiplied by weight, summed and send through a non-linear function

(a) Cell description

Solution: Input - Weight - Sum - Non-linear function - Output

(b) Why do we use non-linear function?

Solution:

- Linear functions can go to ∞
  $\rightarrow$ Human perception has a upper bound. We can experience a maximum amount of pain or happiness. Pain to ∞ means we would die.
  The non-linear function is a foundation of survival

(c) How is learning performed in a perceptron network?

Solution:

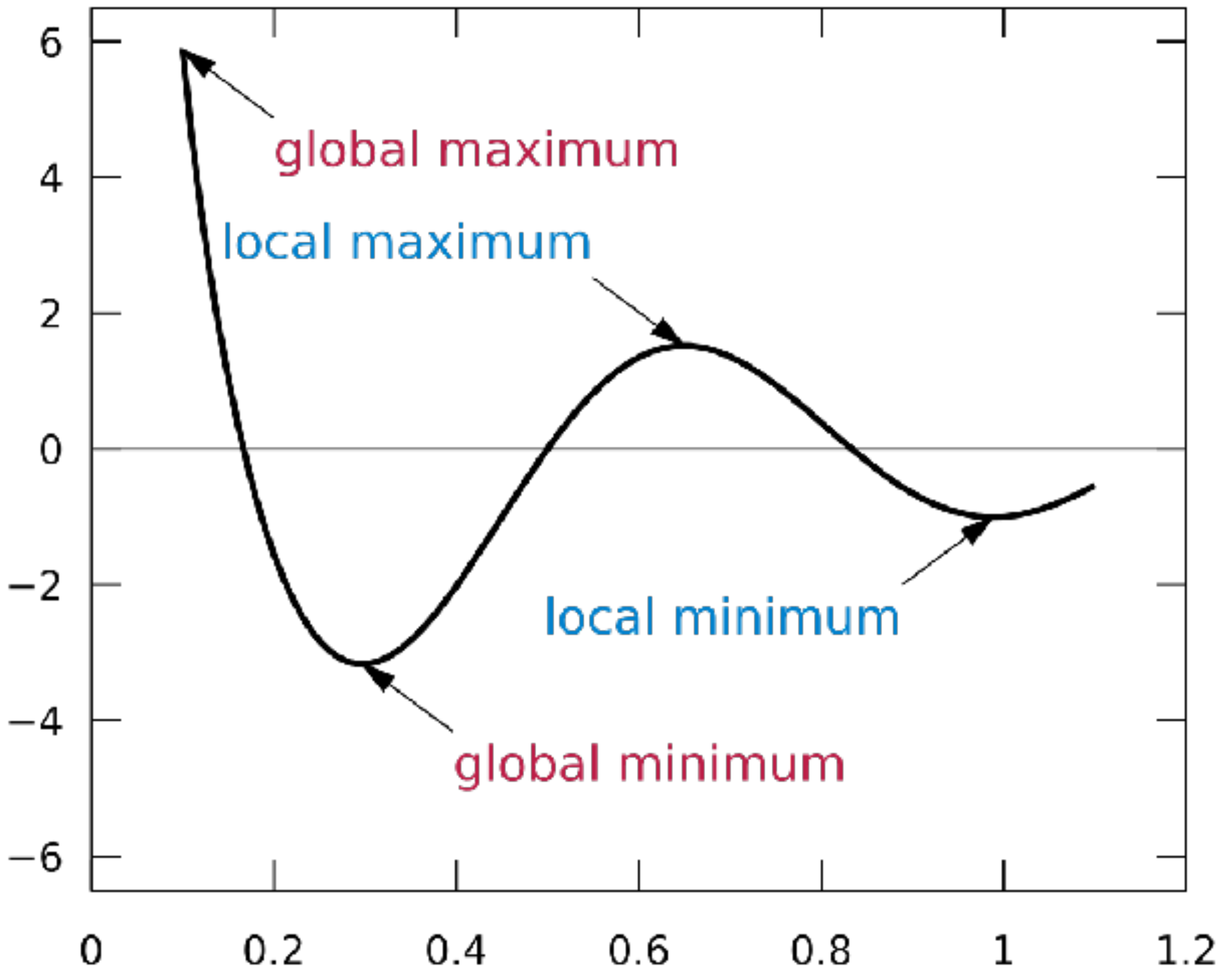
- Initial phase: Start with random weights on the inputs
- Learning phase: Change weights in order to minimize error function

(d) What does the error function describe?

Solution: The difference between the perceptron model and the real model that is derived from training data. The systems aims to minimize the error function.

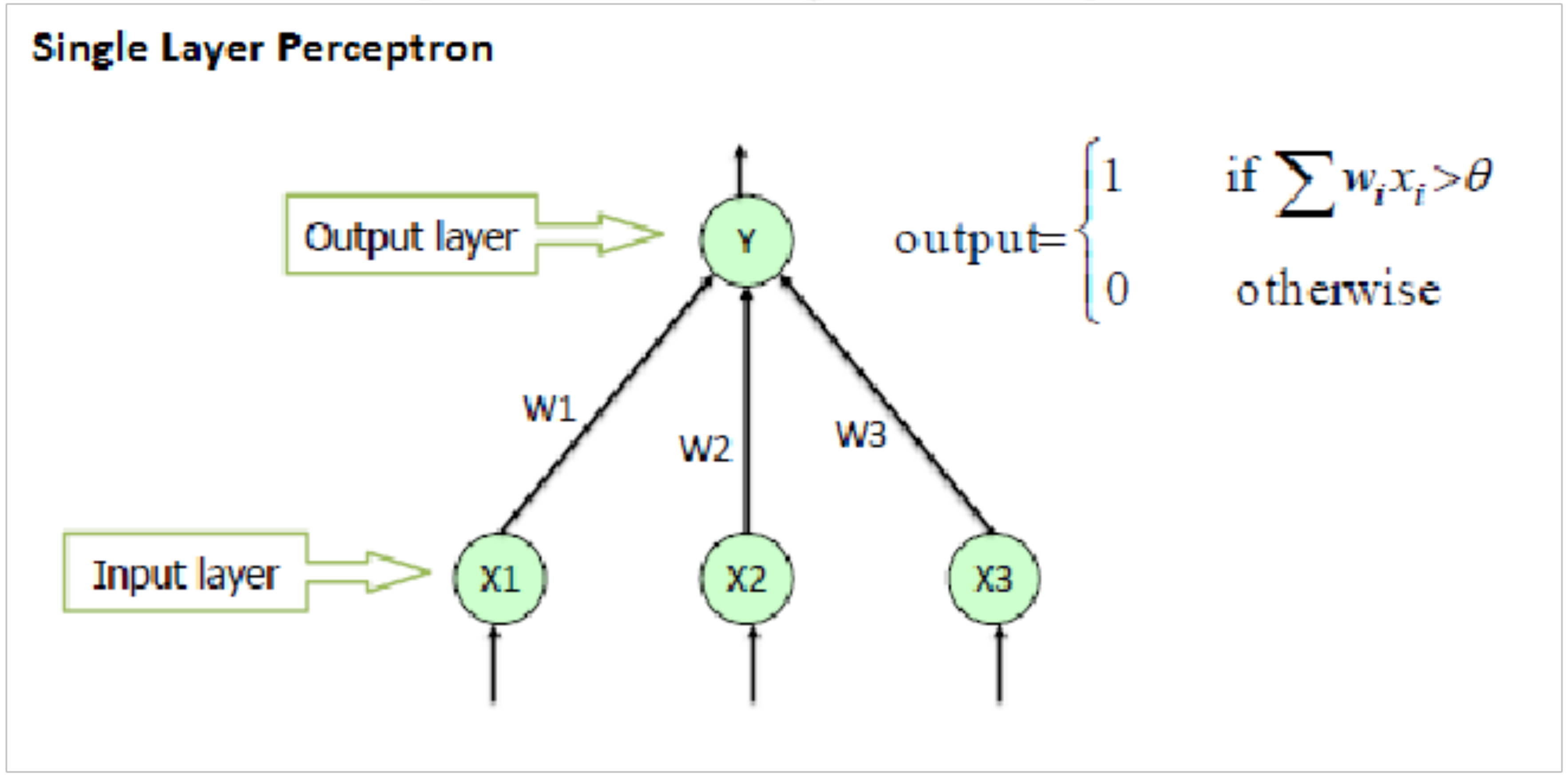
(e) Explain the local minimum problem. Name a method to solve this problem.

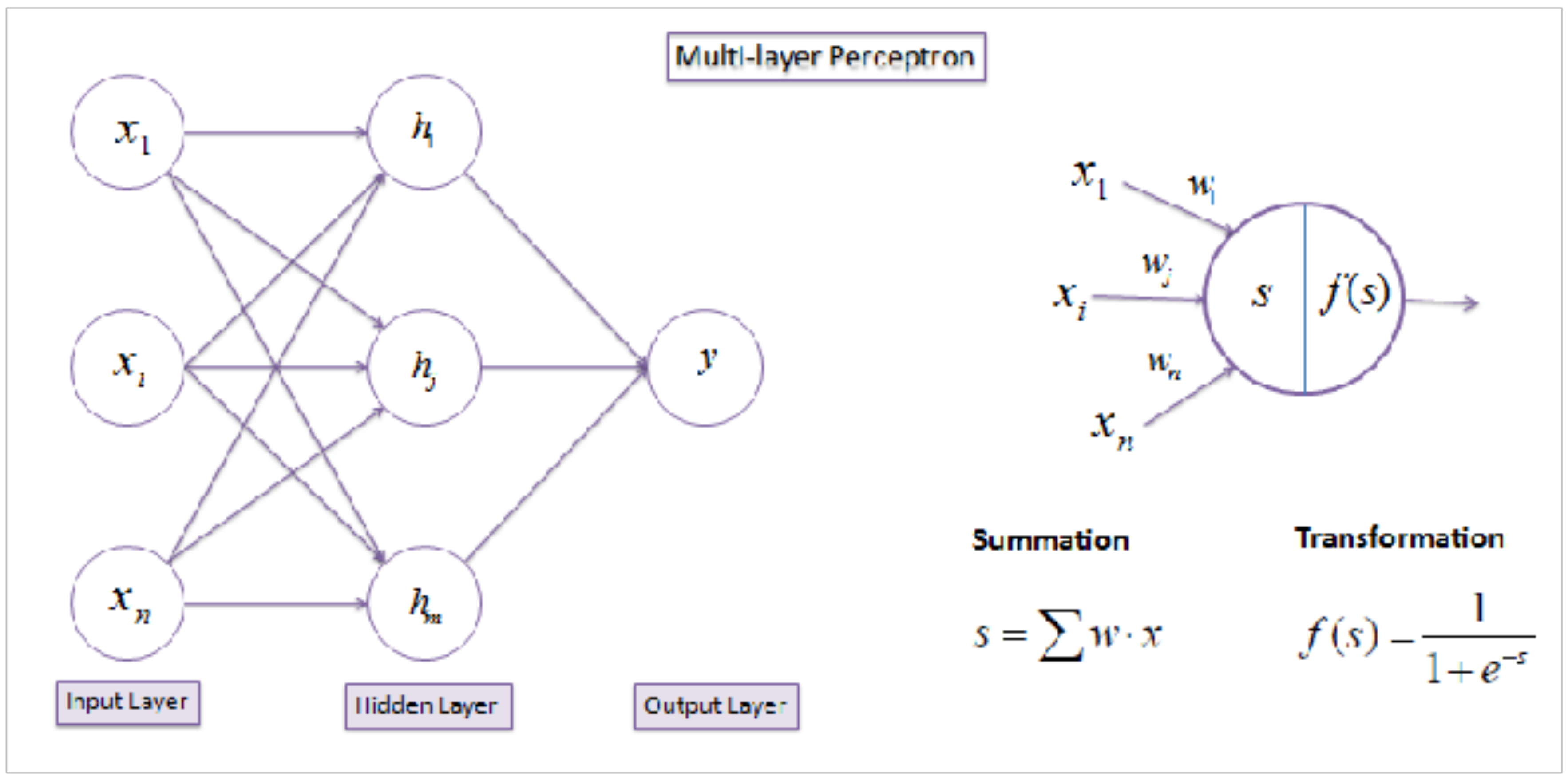
Solution: The local minimum problem describes the finding of a minimum in the error function. Since the error function is unknown to the system it stuck at this point. Possible solutions: Add noise or randomly try other points on the error function (by changing weights).



(f) What is a multilayer perceptron?

Solution: The output of one neuron goes to the input of another neuron.





3 Speech recognition

1. Describe the Acoustic Model

Solution: $p(X|W)$: How likely is an observed sound corresponding to a word sequence

2. Describe the Language Model

Solution: $P(W)$: Describes the likelihood of a spoken word
Dictionaries govern how phones are assembled to form words.
Models:

- Context dependent models
- Cross word models

3. Fundamental equation of ASR

Solution:

$$W' = \text{argmax}_W p(X|W)P(W)$$

Given: observation X
Wanted: corresponding word sequence W
Search: Most likely word sequence W'

4. Explain the advantages of mapping

Solution: There is no need to retrain the whole system. You only need a few speakers of the language for system adaptation.

5. How are speech recognition systems trained and tested?

Solution:

- Speech production as a stochastic process.
- Words, phones, ... → states of the speech production process
- Each state emits observed sounds with a certain probability distribution
- Goal of ASR: find most likely state sequence with discrete or continous HMMs

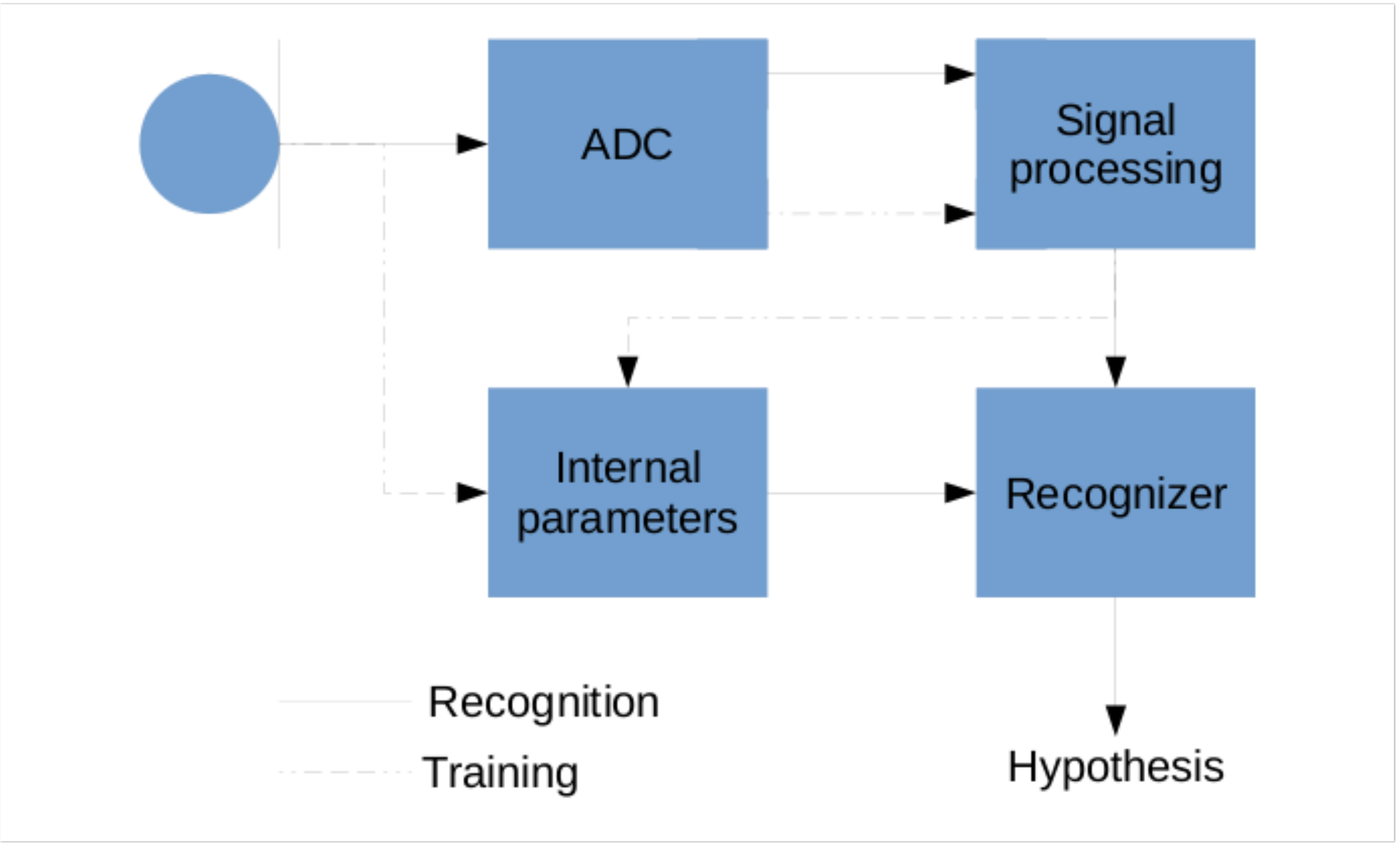
Training

- Introduce pronounciation variants and optional states
- Supervised training method with transcripts that are read to the system

Testing

- Construct big network and prune away unlikely states using the viterbi algorithm

The model is context dependent. It is defined by data driven methods (Decision tree), depending on their phonetic context.



6. Name an evaluation measure and describe its advantages and disadvantages

Solution: Word Error Rate and Word Accuracy

$$\mathrm{WER} = \frac{\#\mathrm{Errors}}{\#\mathrm{Spoken\ Words}}$$
$$\mathrm{WA} = 1 - \mathrm{WER}$$

The WER and WA do not take the word meaning into account. A solution to that problem is to measure the word contribution, for example with the following measures:

- Out of word vocabulary
- Word or slot correct rate

7. Describe the basic principle of speech production

Solution:

- Sounds are un-/ conciousneslytransformed into muscle contraction in the vocal tract
- Ideas are transformed into deformations of the vocal tract as well as pulses are generated by vocal chords
- Speech process order:
  - Vocal chord excitation - vocal tract - speech
  - Excitation: Modeled as sum of sinusoids resulting in speech for quasi stationary frames
  - Vocal tract: Modeled with cubic cylinders

8. What is hearing?

Solution: Henn/Egg problem

- Auditory transduction with the aim to understand content
- Ability to (re)produce heard sounds
- A frequency transformation close to the FFT method (decomposition of complex signals into their main frequency components) is performed in the cochlea.

9. Describe the difference between Hidden Markov Model and Markov Model

Solution:

- Markov Model: Any event can happen after any other event with a certain probabilty
- Hidden Markov Model: The current state is invisible but not its output. It is unknown what happens now and what happens next.

10. What are the knowledge sources (models) of speech recognition? Name 4 variabilities in human speech recognition!

Solution:

- Language model
  - Classic N-Gram models
    Stochastic Language model using local information to predict the likelihood of the next word during search.
  - Context Free Grammars
    Context-free grammars (CFGs) are used to describe context-free languages. A context-free grammar is a set of recursive rules used to generate patterns of strings. A context-free grammar can describe all regular languages and more, but they cannot describe all possible languages.
  - Store HMMs or lattices (word graphs) and use them during decoding
- Acoustic model

Sources of variability & complexity

- Complexity: 50 phonemes, 5000 sounds, 10000+ words
- Variabiltity: Anatomy, speed, loudness, stress, mood
- Ambiguity: Homophones, semantics. Eg:
  "This machine can recognize speech" vs "This machine can wreck a nice beach"
- Segmentation: Continous flow of samples, changing background noise, segementation at utterance level, word level, phone level