1 Detail question

1.1 Phoneme

The base unit of language and refers to the smallest unit which distinguishes between meanings.

1.2 Phonation

refers to the physiological , physical, and neurological processes in the production of a single speech sound, whereby it is a physiological base unit.

1.3 Phone

a specific sound irrespective of its grammatical position or meaning. It is thus aperceptual and acoustic base unit.

1.4 Vocal tract

the cavity in human beings and in animals where the sound produced at the sound source is filtered. (from Wikipedia)

1.5 Formant

The resonances of the vocal tract are known as formants.
They are visible as peaks in their spectral envelopes

1.6 Coarticulation

refers to a situation in which a conceptually isolated speech sound is influenced by, and becomes more like, a preceding or following speech sound.

1.7 Onset

A speech onset is the event when a phonation begins.

1.8 Offset

A speech offset is the ending event of a phonation.

1.9 Intonation

intonation is variation in spoken pitch when used. (from Wikipedia)

1.10 Perceptual modelling

"I don't know" :D

1.11 Source modelling

The purpose of source modelling is to use information available from before to make processing more efficient and improve quality.

1.12 Source-filter model

The source-filter model is a speech source model.
It is loosely based on physiology.
Modelling those characteristic features which have been found to be important for perceptual quality.

1.13 Objective Evaluation Methods

Faster and cheaper than the subjective models.
You can run the tests a lot faster.
If you run an objective test twice, it will give the same result.

Objective evaluation methods:

- POLQA for speech and telecommunications
- PEAQ if music is involved
- PESQ mostly replaced by POLQA

1.14 Subjective Evaluation Methods

The aims of subjective evaluation methods include:

- to quantify preferences of people
- in a setting which reflects actual usage scenarios accurately
- in a structured way such that the test is reproducible (trustworthy)
- that it gives accurate results (useful)
- minimize costs

Subjective evaluation methods:

- P.800 for nave listeners
- P.805 like P.800 but including conversation
- P.835 like P.800 but including noise evaluation
- MUSHRA for experts (10 listeners)

1.15 Fundamental frequency and pitch

The lowest frequency of a periodic waveform. In music, the fundamental frequency is also known as a pitch. (from Wikipedia)

1.16 Concatenative synthesis

Pre-recorded segments of speech are fused into a sequence to obtain the desired word.

1.17 Synthesis by speech production

Formant synthesis or physical modelling synthesis where mechanisms and acoustics of the speech production system is modelled to simulate a real speaker.

1.18 Features

Characteristics which differentiate between speech and noise are called features.
Speech features vary rapidly and frequently.
Some useful features:

- Signal energy
- Estimators for spectral tilt

2 Basic qustions

2.1 Explain how humans produce speech (what physiological processes are involved 3p, what acoustic effect do these processes have 1p and what type of phonations are these related to 2p)

- A phonation begins on a neurological level with the decision or intent to produce speech, whereby the brain sends a message to the physiological organs to produce speech.
- The physiological process begins in the lungs, which contract, increasing the air pressure such that air flows out.
- The acoustical signal is then produced with two mostly independent processes:
  - **voiced** phones are produced by tightening the vocal folds to an appropriate tension, such that they begin to oscillate in the air flow. The varying airflow causes a pressure waveform, that is, a sound.
  - **unvoiced** phones are produced by constricting some part of the vocal tract such that airflow is either prevented or constricted, causing a turbulent mode of airflow and pressure waveform.

2.2 Describe the source filter model of speech production (model description 3p, connection to speech production 2p, application in speech processing 1p)

We use information available from before to make processing more efficient and improve quality.
Speech signals originate from the speech production organs.
Modelling the speech production system can be used to improve speech processing methods.
An accurate physiological model is usually not necessary some very simple approximations can already give most of the benefit.
The source-filter model is one of the most famous source models in speech processing.
In effect, the source-filter model captures following spectral features:

- Envelope shape, or the macro-shape of the spectrum, modelled by the linear predictor
- Fundamental frequency, or the comb-filter shape under the envelope, modelled by the fundamental frequency model

- Harmonics-to-noise ratio (HNR) and overall energy, modelled by the gains of the pitch and noise parts

It is used in speech synthesis and speech analysis and is related to the linear prediction.

2.3 Describe the types of information a speech signal contains

Just a guess:
It contains the fundamental frequency, sampling rate...

3 Algorithmic questions

3.1 Linear prediction

Aims to predict the next sample ϵ_n

$\hat{\epsilon_n} := \sum_{k=1}^{M} \alpha_k \epsilon_{n-k}$ where α_k are weights.

To find the best prediction we need to find the best α_k.

To achieve this we use minimum mean square error:

$min_{a_k} E[|\epsilon_n - \hat{\epsilon_n}|^2]$

Linear prediction is a model of the short-term temporal structure of speech.
Equivalently, it is a model of the spectral envelope.
The most important use of linear prediction is coding for transmission applications (Linear predictive coding).
Parameters of the predictor can be calculated with complexity $O(M^2)$ from the autocovariance, where M is the model order (typically M is in the range 10 to 20).

3.2 Spectral subtraction

It is a noise attenuation method.

Our signal model is $X(z) = S(z) + V(z)$

Given the noise estimate $|\hat{V}(z)|^2$ and the observation $X(z)$, our tank is to estimate speech signal $S(z)$.
The equation for spectral subtraction is:

$$\hat{S}(z) = X(z)\sqrt{\frac{|X(z)|^2 - |\hat{V}(z)|^2}{|X(z)|^2}}$$

3.3 Beamforming

Beamforming refers to methods which use spatial information to extract a specific source from a sound scene.
It uses time-differences between microphones to obtain a better estimate of the desired signal when the source location is known.

- If microphones are at different distances from a source, then the sound will arrive at different times to to microphones.
- Delaying microphone signals appropriately will make the desired source have the same phase in all channels, while other sources are (hopefully) out of phase.
- We can add the delayed sensors whereby in-phase components add up and out-of-phase components attenuate each other.

Types of beamforming:

- Delay-and-sum
- MVDR

3.4 Concatenative synthesis

In concatenative synthesis, segments of speech recordings are copy-pasted to form the desired utterances. It gives best quality, but requires a lot of work and resources.
Bassic approach:

- Collect a database of speech segments with different phonemes.
- Concatenate segments to obtain desired word.
- It is then necessary to overlap-add subsequent segments.

The best quality is achieved with triphones.
Large portion of those triphones rarely occur so they are not used.
We can reduce the set with a language model.

Drawbacks of this approach:

- Someone has to sit down and speak the triphones
- It takes a long time
- The pitch also need to be reproduced

Therefore, the construction of the database is hard.
We can use an existing corpora

3.5 Algebraic codebook

The algebraic codebook generates vectors with an algorithm such that there is no storage required.

3.6 Short-time Fourier transform

Short-time Fourier transform is the most common speech analysis method.

- One window gives a snapshot image of the signal
- Analysis of multiple, consecutive windows gives a movie
- We can analyze how signal is changing over time
- We use a sliding window defined as $\hat{X}_{n,k} = W_n - kX_n$ where each value of k gives a different snaphot of the signal
- Take DFT of each window

Calculation of STFT:

- At position k, apply windowing (typically, Hamming windowing) to obtain segment of the signal of length N.
- Apply the fast Fourier transform to obtain the spectrum $X_k(\omega)$
- Take the logarithm of the absolute value $20 \log_{10} |X_k(\omega)|$ to obtain the logarithmic spectrum
- Advance position by K, that is, k := k + K and return to 1.

3.7 Overlap-add

Overlap-add is a method for windowing a signal such that we can modify the segments and reconstruct the modified signal.
Algorithm:

- Applying windowing function W_n
- Modify/process window with your-algorithm-of-choice.
- Applying windowing function W_n again
- Add overlapping segments together to obtain output signal

Usually we would perform a time-frequency transform on the windowed signal $W_n X_n$ and perform modifications in the frequency-domain.
Almost all frequency-domain processing algorithms are based on overlap-add.

When the reconstructed signal is equal to the original signal, then we have a perfect reconstruction.
Perfect reconstruction works as long as the Princen-Bradley condition holds:

$$\overset{T}{P}L P_L + \overset{T}{P}R P_R = I$$

3.8 Entropy coding

It is a frequency domain coding.
It reduces the bit-rate.
The average bit-rate is 1.5 bits/symbol which is known as Huffman coding.

3.9 Voice activity detection

Refers to a class of methods which detect whether a sound signal contains speech or not.
In a noise-free scenatio this task is trivial but that rarely happens in reality.
The basic idea of the algorithm is:

- Calculate a set of features from the signal which are designed to analyze properties which differentiate speech and non-speech.
- Merge the information from the features in a classifier, which returns the likelihood that the signal is speech
- Threshold the classifier output to determine whether the signal is speech or not.

It is used as a low-complexity pre-processing method

3.10 Fundamental frequency estimation

The fundamental frequency describes a basic property of speech whereby its estimation is perceptually important.
F_0 is visible and can be estimated in many different domains:

- Correlation-analysis in time-domain and autocorrelations show peaks at the distance of the pitch lag and its multiples.
- Magnitude spectra show a comb-structure at the fundamental frequency distance.
- Cepstra show peaks at the distance of the pitch lag and weak peaks also at its multiples.

3.11 Signal-to-noise ratio

The signal to noise ratio is a generic measure for signal distortion and noise.
It is calculated the following way:

$$SNR_{out} = \frac{|S(z)|^2}{|S(z)-\hat{S}(z)|^2}$$

The SNR does not however differentiate between different types of effects.

3.12 Speech distortion index

To quantify how much a method distorts the desired speech signal we can measure how much filtering modifies a clean signal.
Speech distortion index is calculated the following way:

$$SDI = \frac{|A(z)S(z)-S(z)|^2}{|S(z)|^2}$$

3.13 Noise reduction factor

we can quantify how much noise V(z) is attenuated by a filter A(z) using the noise reduction factor.

$$NRF = \frac{|V(z)|^2}{|A(z)V(z)|^2}$$

In many cases it is more important to preserve the original speech signal (keep SDI low) than to maximize NRF, because human listeners find distortions annoying.

For a speech recognition application it might though be more important to obtain the best SNR even at the cost of higher SDI, because (or if) all noise and distortions reduce the accuracy of the speech recognizer equally.

3.14 Mel-frequency cepstral coefficients

Mel-Frequency Cepstral Coefficients is a representation which contains information of the envelope shape of speech signals.

- It takes the logarithm of frequency components, which corresponds to perceptual power-sensitivity.
- It uses the mel-scale to mimic perceptual sensitivity for different frequency regions.
- In addition, it uses DCT or FFT to decorrelate the down-sampled mel-frequency spectrum.
- Computationally simple to implement

There are also some drawback:

- Poor performance in noisy conditions
- Choice of smoothing filter is arbitrary
- Useful for analysis only

3.15 Zero-crossing rate

It counts the number of times the signal crosses zero (=changes sign) within a window.
It's very simple to implement.
A low-frequency signal will cross zero only sometimes.
A high-frequency signal will cross zero all the time.

3.16 Cepstrum

Taking the Fourier transform of the log-spectrum is know as cepstrum.
The x-axis of the cepstrum is known as the quefrency axis and it is a time-domain.
Filtering in the cepstrum domain is known as liftering.

3.17 Pulse-code modulation

PCM most common (high-quality) storage format for digital speech and audio signals.
Variations of PCM:

- The sampling rate
- Quantization algorithm

3.18 Unoform quantization

In uniform quantization, the signal s_k is rounded to the nearest quantization level and the error made is independent of the signal magnitude.

3.19 Logarithmic quantization

In logarithmic quantization the quantization error is relative to the magnitude.
Here for a signal s_k we first take the logarithm of the magnitude log, then quantize round, return to the linear scale exp and restore the sign signal.

3.20 mu-law quantization

The μ-law rule is an approximation of logarithmic quantization which avoids this problem.

4 Discussion

4.1 Codecs

AMR is the most successful codec and is still widely used.

Linear prediction tries to model the spectral envelope.
Line Spectral Frequencies is the most common/effective representation of linear predictors for quantisation.

Codecs also try to model the fundamental frequency.
Long time prediction algorithm is used to achieve that.
LTP is a vector codebook and is signal adaptive.

After we have captures the spectral envelope with linear predictor and the fundamental frequency with the long time prediction, we are left with a residual which is practically a noise.
Residual coding aims to capture that noise.

- We first encode the gain (energy) of the noise vector
- we encode the fixed-length residual with an algebraic codebook
- The algebraic codebook generates vectors with an algorithm such that there is no storage required
- The best quantization is found by a brute-force search, also known as the analysis by synthesis method

4.2 Enhancement

Single channel

- SPECTRAL SUBTRACTION (explained above)
- WIENER FILTERING defined as: $A(z) = \frac{|X(z)|^2 - |\hat{V}|^2}{|X(z)|^2}$

- VOICE ACTIVATION DETECTION
- LINEAR FILTER

There are different types of performance measures.
For example we can listen to the output in a same environment and hardware as the intended application.
We can also listen to a whole range of speakers (male, female, child, etc).

We can also use more technical approaches like: signal to noise ratio, speech distortion index and noise redution factor.

Multi channel

- Beanforming (EXPLAINED ABOVE)
- Dereverberation methods, which attempt to reverse the effect of room-acoustics on the desired speech signal