

# Speaker and Speech Recognition

## 1 Questions - Speaker Recognition

1. Specify the four criteria by which the variants of speaker recognition can be differentiated!
2. What is the meaning of discriminative training?
3. Give an overview of the preprocessing for speaker recognition (slide 14). When used for segmentation, the attenuation of the noise suppression is not limited to a lower bound (slide 15). Why is that?
4. Which features are used for speaker recognition? Compare the feature extraction methods of speech recognition and speaker recognition.
5. What do the attributes "open" and "closed" of the Wiener filter mean in a mathematical way? How do the filter coefficients  $H$  look like, respectively? How can the classification ("open" / "closed") be used as preprocessing for speaker recognition?
6. What is a (speaker specific) threshold codebook and why is it used?
7. Explain the approach to use GMMs for speaker recognition (slide 24). What is denoted as  $\mathbf{X}$ ,  $\mathbf{g}$ ,  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ ?
8. Compare the evaluation results of speaker recognition based on codebooks and GMMs!

## 2 Questions - Speech Recognition

1. Which circumstances complicate the automatic recognition of spoken language?
2. Which variants of speech recognition systems can be distinguished? Try to find application examples for some of the variants.
3. Which characteristics of a speech recognition system can be used for its evaluation? How can word error rates larger than 100% be archived?
4. What is the statement of Bayes' theorem. How is it applied to speech recognition?
5. What is a semi continuous HMM? How can memory and computational load be reduced using semi continuous HMMs?
6. What is acoustic model generation? What is a language model? Which operations are being executed at the training and which at runtime of the speech recognition system?

7. How can a continuous speech recognizer be generated on the basis of HMMs of elementary units?
8. Finally, give an overview of speech recognition on the basis of slide 42.

### 3 Answers - Speaker Recognition

1. See slides 6 – 10: Differentiation of a) verification versus identification, b) text dependent versus text independent, c) "closed" versus "open" identification, and d) discriminative versus non discriminative training.
2. See slide 10. (More details can be found on slide 36.)
3. See slide 14. When used as a voice activity detection (VAD), an attenuation limit ( $H_{\min}$ ) of the Wiener filter would produce an offset in the filter opening constant (denoted with  $0.1 \dots 0.3$ ), which is not helpful. The attenuation limit is just used in noise reduction to prevent musical noise.
4. For speaker recognition, phoneme independent speaker dependent features are required, while for speech recognition, phoneme dependent speaker independent features should be used. Nevertheless for both, cepstral coefficients and MFCCs are being used. See also slides 17 and 18.
5. See slide 15: The filter is called to be open if  $H$  is above a certain threshold (e.g.,  $0.1 \dots 0.3$ ). This classification can be used as a simple voice activity detection (VAD), in order to use only those segments that contain speech for speaker recognition.
6. A threshold codebook contains individual thresholds for the distances between code vectors and test feature vectors. Based on this criterion, the speaker will be accepted or refused.
7. The probability that the  $GMM(\mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$  produces the observed series of feature vectors  $\mathbf{X}$  is to be calculated. This is done both for a speaker model ( $s$ ) and a background model ( $b$ ). The matrix  $\mathbf{X}$  contains the observed feature vectors of the active segments of an utterance.  $\mathbf{g}$  are the weights of the Gaussian curves of the GMM.  $\boldsymbol{\mu}$  are the mean values of the Gaussian curves.  $\boldsymbol{\Sigma}$  is the covariance matrix.
8. See slides 27 to 29.

### 4 Answers - Speech Recognition

1. Spoken language is individual, dialect dependent, spontaneous, perhaps grammatically not correct, and continuous (word boundaries cannot be recognized); furthermore, the microphone signal may contain background noise, etc.
2. Single word recognition / keyword spotting / connected words recognition. speaker dependent / speaker independent / speaker adaptive; size of the vocabulary;
3. The word error rate in relation to the target scenario (e.g., SNR), hardware requirements (e.g., PC or DSP), real time capability, .... In theory, the "number of insertions" could be higher than the number of words to be recognized and thus also the word error rate can get arbitrarily high (see equation on slide 46).

4. In a nutshell, using the Bayes' theorem, "condition" and "event" in conditional probabilities can be inverted,  $P(A|B) = P(B|A) P(A)/P(B)$ . In speech recognition, the starting point is the probability that a word sequence  $W$  is being spoken, assuming a given feature sequence  $X$ . But in the decoding problem, the probability that a given word sequence  $W$  "generates" a certain feature sequence  $X$  is calculated. The conversion is exactly the statement of Bayes' theorem.
5. See slide 51 + 52.
6. The acoustic model describes the connection between feature sequences  $X$  and word sequences  $W$ . The model generation is described on slide 53. The language model gives estimations for the (a-priori) probability of a word sequence  $W$ . At runtime: Feature extraction, decoding (sometimes also adaption). At training time: Feature extraction, text processing, and training of both the acoustic model and the language model (see slide 55).
7. This can be done using parallel and serial merging of HMMs. Parallel merging needs a parameter  $\lambda$  for the probabilities that define which HMM will be traversed. (This can be derived from the language model.)