Speech Signal Processing (SSV)

Examiners	Prof. Dr. Timo Gerkmann
Exam date	09/16/2021
Department	Informatik
Labels	Zweittermin mit Lösung mündlich

This is a memory protocol of my oral exam in speech signal processing, held by Prof. Dr. Timo Gerkmann in the summer term 2021. Since this is a memory protocol, it cannot be guaranteed that the protocoll is complete or correct.

The oral exam was held over Zoom. In addition to a normal video conference setup, a second device to capture handwritten documents was needed (either a tablet to write on or a smartphone facing the sheets of paper to write on).

In general, the exam was very similar to the previous protocols. There were three parts to the exam: the source-filter model, quantization, and speech enhancement. The actual questions in the exam were not formulated like I did in this protocol, it was more of a relaxed conversation. The exam was in English and took about 35 minutes. In addition to me and the professor, a research assistant was also present. I received a grade of 1.0.

Source-Filter model

- Sketch the source-filter model of speech production
  - I sketched the model with the generation of the unvoiced and voiced (with f_0) excitation signal, the V/UV decision and the vocal tract filter.
  - I added $e(t)$, $h(t)$, and $s(t)$ to denote the excitation signal, the vocal tract filter, and the resulting speech signal.
  - I wrote the equation $s = e * h$ for the signal, he asked in what domain the signal is defined and I changed the equation to $s(t) = e(t) * h(t)$.
- How would this formula look in the frequency domain?
  - I did the z-transform, resulting in $S(z) = E(z) \cdot H(z)$
- For the convolution, we would need an infinitely long impulse response which we don't have in practice. What would we use instead?
  - I explained that the ARMA-model would be used and noted the equation: $s(t) = b_0 e(t) - \sum_{k=1}^{\nu} a_k s(t-k)$
  - I explained the moving-average part and why it is not needed (except for b_0) as well as the autorecursive part.
  - He asked if t refers to time or is an index, I said it is an index (I guess it would have been clear if I used n instead.
- How many coefficients a_k should be used?

    - I said that for a sampling frequency of 8kHz, which is typical for telephony, there would be a maximum frequency of 4kHz (due to the sampling theorem). We expect one resonance (i.e., formant) per kHz and for each resonance, two coefficients are required. In addition, two coefficients are needed to avoid aliasing effects. Therefore, I suggested to use 10 coefficients
- What do you think is the sampling rate used in Zoom?
    - I said that it would be more than 8kHz (telephony) and less than 44kHz (CD), probably around 16kHz (HD voice) to 32kHz (HD voice +)
- Then, I was asked to add b_0 to my sketch of the source-filter model
    - I added it as a multiplication before the filter h.
- How would the spectrum vor a voiced sound look like?
    - I sketched a spectrum and pointed out formants and the fundamental frequency

Quantization

- What is quantization?
    - I said that as sampling is discretization in time, quantization is a discretization of values because computers cannot store infinitely precise numbers.
- What types of quantization do you know?
    - I started with the uniform quantizer and explained that the quantizer quantizes the signal with a fixed step size. To explain that, I drew the step-like function mapping continuos values to their quantized values. He then asked me if I drew a mid-rise or mid-tread function.
    - Then, he asked how the noise of the quantizer would be described. I mentioned the signal to noise ration and the noise power. He wanted to see the equation for noise power: $P_N = \frac{\Delta x^2}{12}$
    - Then, I said that there are also non-uniform quantizers that use compressors (I mentioned A-law and μ-law) to improve the quantization. He asked me why this would be necessary and I answered that speech is not uniformly distributed and that there are many values close to zero and fewer extreme values. I was then asked to draw a typical distribution of speech. I did that and mentioned that it is typically a supergaussian distribution.
    - Finally, I mentioned that in telephony Adaptive Quantization is used, often in combination with Differential Pulse Code Modulation. I did not remember the exact name, but I could explain what it is and why it is used.

Speech Enhancement

- Assumed there is a signal $Y(f) = S(f) + N(f)$. What would you do to estimate the speech signal?
    - I said that I would use the Wiener filter wrote the equation $\hat{S}(f) = G(f) \cdot Y(f)$ and $G(f) = \frac{\sigma^2_{S,f}}{\sigma^2_{S,f} + \sigma^2_{N,f}}$.
- What is the value range for $G(f)$?
    - Between zero and one.
- When does it take these values?
    - It is zero when there is only noise ($\sigma_S = 0$) and one when there is only noise ($\sigma_N = 0$)
- What is the criterion that is minimized for the filter?
    - The minimum mean squared error criterion: $E((S(f) - \hat{S}(f))^2)$
- And what is it minimized for?
    - $G(f)$
- How would the minimization be done?
    - I would take the first derivative of the error criterion with respect to $G(f)$ and solve for zero.

- If we would minimize for $\hat{S}(f)$ directly instead of $G(f)$, what would be the result?
    - That would be the same as dropping the constraint that the filter should be linear. When we add the constrained that signal and noise are Gaussian dristibuted, we would still get the Wiener filter. Without this constraint, we would get different filters.
- If we had two microphones, how would we reduce noise?
    - The simplest solution would be the delay-and-sum beamformer. I sketched two microphones and a far-field audio source and explained the beamformer.
    - I also said that this beamformer is only useful for uncorrelated noise between the microphones because otherwise, some noise could not be attenuated by the beamformer.
- Could you explain the MVDR beamformer and write the formula down?
    - I wrote the formula $\hat{s} = \frac{a^H \Phi_{NN}^{-1}}{a^H \Phi_{NN}^{-1} a} y$ and explained the meaning of a and Φ_{NN}.
    - He asked what the form of the noise correlation matrix would be: a 2x2 matrix
    - He wanted to know how the steering vector would look like. I said that each element would have a time delay in the form of $z^{-\tau}$. I wanted to write down how it would look in the Fourier domain, but I forgot the correct transformation from z-domain. (It would have been $e^{-j\tau\omega}$.) He also wanted to hear that the first element in the vector would be a factor of 1 (no delay) for the reference microphone.