Selected Topics in Deep Learning for Audio, Speech, and Music Processing

Exam Questions

Dr.-Ing. Stefan Balke, Dr.-Ing. Andreas Brendel, Dr. Ben Maman, Prof. Dr. Meinard Müller
Friedrich-Alexander Universität Erlangen-Nürnberg
International Audio Laboratories Erlangen

02. 11. 2026

First Name: _____ Last Name: _____

Matriculation Number: _____

Study Programme: _____

Exam Date: _____

Exam Duration: _____ min

Important:

- Each question has one final answer: a **single number**, such as the integer 2 or decimal value 1.25. Where needed, the question specifies units and rounding.
- Each correct answer is worth **exactly 1 point**.
- All questions are independent and can be solved by hand. **No digital devices** are needed or permitted.

Best wishes for the exam!

Grade	**1.0**	**1.3**	**1.7**	**2.0**	**2.3**	**2.7**	**3.0**	**3.3**	**3.7**	**4.0**	**Fail**
Points	48–55	45–47	42–44	40–41	37–39	34–36	31–33	28–30	26–27	23–25	0–22

Points: _____/ _____

Grade: _____

Signature of Docent: _____

Unit 1: Introduction to Audio and Speech Processing /22 Points

Question 1.1 An STFT uses sampling rate $F_s = 8000$ Hz, DFT length $N = 4000$, and hop size $H = 2000$ samples. What is the frequency spacing between adjacent DFT bins, in Hz?

2

Question 1.2 An STFT uses sampling rate $F_s = 8000$ Hz, DFT length $N = 4000$, and hop size $H = 2000$ samples. What is the time spacing between adjacent STFT frames, in milliseconds?

250

Question 1.3 A mono speech recording contains 32000 samples at sampling rate $F_s = 16000$ Hz.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: the recording duration is 2 seconds
Label **2**: resampling to 8000 Hz while preserving duration gives 16000 samples
Label **4**: the Nyquist frequency is 16000 Hz
Label **8**: playing the original samples at 8000 Hz would make the recording last 4 seconds

11

Question 1.4 A synthesized trumpet tone has fundamental frequency $F_0 = 200$ Hz. Its pitch is raised by two octaves. What is the new fundamental frequency, in Hz?

800

Question 1.5 A harmonic tone has spectral peaks at 450 Hz and 750 Hz. Its fundamental frequency is known to lie between 100 and 200 Hz, and both peaks are harmonics of that fundamental.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: the fundamental frequency is 150 Hz
Label **2**: the 750 Hz peak is the fifth harmonic
Label **4**: exactly four harmonics lie in the interval from 500 to 1000 Hz, inclusive
Label **8**: 1400 Hz is a harmonic

3

Question 1.6 A synthesized trumpet tone has fundamental frequency $F_0 = 240$ Hz. It is transposed one octave downward. A spectral peak of the transposed tone occurs at 600 Hz. What is the harmonic number of this peak, counting the fundamental as harmonic 1?

5

Question 1.7 A pitch control signal is $F_0(t) = 200 + 5\sin(2\pi \cdot 4t)$ Hz, where t is measured in seconds. What is the maximum fundamental frequency, in Hz?

205

Question 1.8 A piece is played at 120 beats per minute.
What is the duration of one beat, in milliseconds?

500

Question 1.9 A piece is in 3/4 time and played at 30 quarter notes per minute. What is the duration of one measure, in seconds?

6

Question 1.10 During a voiced speech segment, the vocal folds complete 25 vibration cycles in 0.1 seconds. What is the fundamental frequency, in Hz?

250

Question 1.11 A voiced speech signal has fundamental frequency $F_0 = 200$ Hz. What is the duration of one fundamental period, in milliseconds? **5**

Question 1.12 An amplitude control signal is $a(t) = 0.6 + 0.2\sin(2\pi \cdot 5t)$, where t is measured in seconds. What is the minimum amplitude? **0.4**

Question 1.13 A vibrato control signal is $F_0(t) = 300 + 8\sin(2\pi \cdot 5t)$ Hz, where t is measured in seconds.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: the signal completes 10 vibrato cycles in 2 seconds
Label **2**: its maximum fundamental frequency is 310 Hz
Label **4**: the difference between its maximum and minimum frequencies is 16 Hz
Label **8**: the vibrato frequency is 5 Hz **13**

Question 1.14 A synthesized musical signal is $x(t) = \cos(2\pi \cdot 400t) + \cos(2\pi \cdot 600t)$, where t is measured in seconds. What is the smallest positive repetition period of the complete signal, in milliseconds? **5**

Question 1.15 A voiced speech signal has a fundamental period of 8 milliseconds. A pitch-shifting system raises its pitch by one octave while preserving its duration. What is the new fundamental frequency, in Hz? **250**

Question 1.16 In a simplified source-filter model of voiced speech, the excitation has harmonics at integer multiples of $F_0 = 150$ Hz. A vocal-tract resonance amplifies components in the interval from 1000 to 1500 Hz, including the boundaries. How many excitation harmonics lie in this interval? **4**

Question 1.17 A voiced speech spectrum contains peaks at 750, 875, and 1000 Hz. Assume these peaks are three consecutive harmonics of the excitation. What is the fundamental period, in milliseconds? **8**

Question 1.18 A speech-recognition model represents each analysis frame by 40 features. For each prediction, it concatenates the current frame with the four preceding and four following frames. How many entries does the resulting input vector contain? **360**

Question 1.19 An STFT uses sampling rate $F_s = 12000$ Hz, DFT length $N = 8000$, and hop size $H = 6000$ samples.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: the frequency spacing between adjacent DFT bins is 2 Hz
Label **2**: the hop duration is 0.25 seconds
Label **4**: the STFT produces 200 frames per second
Label **8**: the STFT produces 4 frames per second **0**

Question 1.20 For the STFT, add the labels of all true statements (none: **0**; all: **15**).
Label **1**: it applies Fourier analysis to successive signal segments
Label **2**: the hop size is the spacing between consecutive frames
Label **4**: the hop size is always half the window length
Label **8**: a longer DFT gives finer frequency spacing **11**

Question 1.21 Consider the source-filter model of voiced speech.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: the vocal-fold source is periodic
Label **2**: the vocal tract shapes the spectral envelope
Label **4**: formant frequencies must be integer multiples of the fundamental frequency
Label **8**: articulation can change formants while F_0 stays fixed _____ **11**

Question 1.22 For voiced and unvoiced speech, add the labels of all true statements (none: **0**; all: **15**).
Label **1**: voiced speech typically has periodic vocal-fold vibration
Label **2**: F_0 is the number of vocal-fold vibration cycles per second
Label **4**: vocal-tract resonances are called formants
Label **8**: unvoiced speech always has a well-defined F_0 _____ **7**

Unit 2: Automatic Music Transcription /22 Points

Question 2.1 Consider common music representations.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: a waveform directly specifies the notes being played
Label **2**: Mel representations use perceptual frequency scales
Label **4**: an STFT represents a signal over time and frequency
Label **8**: MIDI and piano rolls are symbolic representations _____ **14**

Question 2.2 A music transcriber represents A4 with MIDI number 69. What MIDI number represents the same pitch two octaves higher? _____ **93**

Question 2.3 Assume equal temperament, with A4 at 440 Hz and MIDI number 69.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: A5 has frequency 880 Hz
Label **2**: MIDI number increases linearly with frequency
Label **4**: each semitone adds the same frequency difference in Hertz, regardless of pitch
Label **8**: an octave above A4 has MIDI number 83 _____ **1**

Question 2.4 A frame-level transcription has active notes C4, E4, G4, C5, and E5. Chroma features merge pitches across octaves. How many chroma bins are active? ____ **3**

Question 2.5 A piano-roll excerpt contains A3, C4, A4, C5, E5, and A5. How many pitch classes occur in more than one octave? _____ **2**

Question 2.6 A training set contains 50 excerpts. Pitch-shift augmentation creates 11 shifted copies of each excerpt, in addition to the original. Then each resulting excerpt is time-scaled to create 2 slower and 2 faster versions, in addition to itself. How many excerpts are in the augmented set? _____ **3000**

Question 2.7 A transcription model estimates pitch activity for 88 pitches in each of 50 frames. It stores these estimates in a matrix with one row per pitch and one column per frame.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: one matrix column represents the pitches active at a particular frame
Label **2**: a polyphonic signal must have at least two active pitches in every frame
Label **4**: a monophonic frame can have multiple active pitches at once
Label **8**: the matrix contains 4400 pitch-activity values _____ **9**

Question 2.8 A polyphonic transcriber predicts 0.72 for C4, 0.50 for C$^{\#}$4, 0.51 for D4, 0.49 for E4, and 0.88 for G4. With a strict activation threshold of 0.5, how many pitches are active?
_____ **3**

Question 2.9 An onset posteriorgram has values:

$$P(1) = 0.8,\ P(2) = 0.7,\ P(3) = 0.5,\ P(4) = 0.1,\ P(5) = 0.7,\ P(6) = 0.2,\ P(7) = 0.1.$$

Three C4 onsets are initially assigned to frames 2, 4, and 6. Greedy snapping processes them in time order, searches one frame on either side, and assigns each onset to an available frame with the highest activation.

Add the labels of all true statements (none: **0**; all: **15**).

Label **1**: the first onset is assigned to frame 1

Label **2**: the second onset is assigned to frame 5

Label **4**: the third onset is assigned to frame 7

Label **8**: the activations at the newly assigned frames sum to 1.0 _____ **3**

Question 2.10 A transcription model represents pitches from A0 to C8, inclusive. How many pitches are represented? _____ **88**

Question 2.11 A frame is encoded by a multi-hot vector for all pitches from C4 through B5. Only C4, E4, and G4 are active.

Add the labels of all true statements (none: **0**; all: **15**).

Label **1**: the vector has 24 entries

Label **2**: the vector contains three entries equal to 1

Label **4**: each entry represents one pitch

Label **8**: multiple entries can equal 1 when multiple pitches are active _____ **15**

Question 2.12 An onset prediction is correct if its pitch matches a reference note and its onset differs by at most 50 ms. The reference notes are C4 at onset 36.70 s and E4 at 40.00 s.

Add the labels of all true statements (none: **0**; all: **15**).

Label **1**: predicting C4 at 36.74 s is correct

Label **2**: predicting D4 at 36.72 s is correct

Label **4**: predicting E4 at 40.07 s is correct

Label **8**: predicting F4 at 40.07 s is correct _____ **1**

Question 2.13 A pitch-only transcription model represents pitches from C2 to B6. A frame has C4, G4, A6, and B6 active. The audio is shifted up two semitones, and pitches outside the model's range are discarded.

Add the labels of all true statements (none: **0**; all: **15**).

Label **1**: C4 shifts to D4

Label **2**: G4 shifts to B4

Label **4**: A6 shifts to C7

Label **8**: exactly one active pitch shifts outside the model's range _____ **9**

Question 2.14 A frame-level pitch transcription gives C4 activity $[0, 1, 1, 0, 1]$ and E4 activity $[0, 0, 1, 1, 0]$ across five consecutive frames.

Add the labels of all true statements (none: **0**; all: **15**).

Label **1**: C4 is active in three frames

Label **2**: both C4 and E4 are active in the third frame

Label **4**: every frame has exactly one active pitch

Label **8**: exactly one frame is silent _____ **11**

Question 2.15 An audio recording is paired with a digitized score.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: the score gives note pitches and musical positions
Label **2**: EM alternates between alignment and transcriber training
Label **4**: the score notes are assumed to occur in the recording
Label **8**: alignment is generally easier than transcription _____ **15**

Question 2.16 Consider onset snapping after sequence alignment.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: note durations and offsets remain difficult transcription problems
Label **2**: sequence-level DTW alignment guarantees accurate note offsets
Label **4**: onset snapping adjusts estimated onset times toward strong onset activations
Label **8**: greedy and graph-based methods are possible snapping strategies ____ **13**

Question 2.17 A transcription model predicts activity for P pitches over T time frames.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: its frame-level output can be arranged as a $P \times T$ matrix
Label **2**: monophonic pitch prediction can be treated as multiclass classification
Label **4**: polyphonic pitch prediction can be treated as multiple binary classifications
Label **8**: a polyphonic frame can be silent _____ **15**

Question 2.18 Consider the Onsets and Frames piano transcription model.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: the model uses convolutional and LSTM layers
Label **2**: frame-level activity alone always gives precise note offsets
Label **4**: onset predictions indicate note beginnings
Label **8**: frame predictions estimate pitch activity over time _____ **13**

Question 2.19 Consider piano transcription data recorded with a Disklavier.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: the paired MIDI has no timing information
Label **2**: the aligned MIDI can provide note-level transcription targets
Label **4**: recordings can be paired with MIDI performance data
Label **8**: the audio and MIDI are nearly perfectly aligned _____ **14**

Question 2.20 Consider common transcription evaluation metrics.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: precision is true positives divided by predicted positives
Label **2**: recall is true positives divided by reference positives
Label **4**: F1 is the harmonic mean of precision and recall
Label **8**: F1 can never be zero _____ **7**

Question 2.21 Consider the value ranges of precision, recall, and F1.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: precision, recall, and F1, when defined, each lie between 0 and 1
Label **2**: F1 is always greater than both precision and recall
Label **4**: F1 lies between precision and recall
Label **8**: precision, recall, and F1 can exceed 1 _____ **5**

Question 2.22 Consider automatic music transcription with P pitches and I instruments. Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: instrument-agnostic transcription predicts pitch only
Label **2**: instrument-aware transcription predicts pitch and instrument
Label **4**: a joint model needs $P + I$ classes for P pitches and I instruments
Label **8**: a joint model needs $P \times I$ classes for P pitches and I instruments — **11**

Unit 3: Nonnegative Music Audio Decomposition /15 Points

Question 3.1 For music source separation from mono or stereo recordings:
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: the goal is to separate sources from an audio mixture
Label **2**: the number of sources can exceed the number of input channels
Label **4**: musical sources can be highly dependent
Label **8**: each source requires its own input channel

7

Question 3.2 NMF represents a spectrogram with 100 frequency bins and 200 time frames using 5 spectral templates. How many values are stored in the templates and their framewise activations altogether?

1500

Question 3.3 Consider nonnegative matrix factorization (NMF).
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: jointly optimizing the template and activation matrices is a convex problem
Label **2**: NMF constrains its factors to be nonnegative
Label **4**: nonnegativity can encourage a semantically meaningful decomposition
Label **8**: nonnegativity prevents template vectors from canceling one another

14

Question 3.4 Consider constraints in NMF-based audio decomposition. Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: under multiplicative updates, entries initialized to zero remain zero
Label **2**: a synchronized score can constrain templates but not activations
Label **4**: zero-valued template entries can become positive during multiplicative updates
Label **8**: harmonic template constraints alone ensure sharp note onsets

1

Question 3.5 For one audio frame, let $V = \binom{4}{3}$, $W = \left(\begin{smallmatrix}1 & 2\\ 2 & 1\end{smallmatrix}\right)$, and $H = \binom{2}{2}$. What is the squared reconstruction loss $L = \sum_{k=1}^{2}(V_k - (WH)_k)^2$?

13

Question 3.6 A single-template NMF model uses $w = \binom{1}{2}$ and an observed frame $v = \binom{3}{4}$. Its activation is currently $h = 1$. The multiplicative update is $h_{\text{new}} = h\frac{w^\top v}{(w^\top w)h}$. What is the updated activation?

2.2

Question 3.7 An NMF activation vector is initialized as $h = (h_1, h_2, h_3) = (2, 1, 0)$. In every iteration, the entries are updated simultaneously as $h_1 \leftarrow 0.5h_1$, $h_2 \leftarrow 2h_2$, and $h_3 \leftarrow 10h_3$. After three iterations, what is the difference between the largest and second-largest activation entries?

7.75

Question 3.8 An NMF template for $F_0 = 200$ Hz has entries at $100, 200, \ldots, 900$ Hz, spaced 100 Hz apart. Under multiplicative updates, how many entries must be initialized to zero to allow only harmonics of F_0?

5

Question 3.9 A nonnegative autoencoder receives $v = \binom{2}{1}$. Its encoder computes $h = \max\left(\left(\begin{smallmatrix}1 & -3\\ 2 & 1\end{smallmatrix}\right)v, 0\right)$, and its decoder computes $\hat{v} = \max\left(\left(\begin{smallmatrix}1 & 2\\ 3 & 1\end{smallmatrix}\right)h, 0\right)$, with both maxima applied entrywise. What is the first entry of $\hat{v}$?

10

Question 3.10 A score-informed autoencoder produces code $h = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$. A binary score mask $m = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$ gives $h' = h \odot m$, where $\odot$ denotes entrywise multiplication. One frequency bin has decoder weights $(1, 2, 3)$. What is its reconstructed magnitude $(1, 2, 3)h'$? **14**

Question 3.11 A nonnegative autoencoder updates a weight by $w_{\text{new}} = \max(w - \eta g, 0)$. For $w = 0.3$ and gradient $g = 2$, what is the smallest learning rate η that makes $w_{\text{new}} = 0$? **0.15**

Question 3.12 Consider a nonnegative autoencoder (NAE). Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: the encoder maps an input to a lower-dimensional code
Label **2**: the decoder reconstructs the input from the code
Label **4**: ReLU activations can make the code and reconstruction nonnegative
Label **8**: projected gradient descent can keep learned weights nonnegative **15**

Question 3.13 An autoencoder processes a spectrogram with 100 frequency bins and 500 time frames. Each frame is mapped to a code of dimension 10 by a fully connected encoder and reconstructed by a fully connected decoder. Both layers have no biases, and their weight matrices are learned independently. How many scalar parameters are learned in total? **2000**

Question 3.14 In audio mosaicing, the fixed source spectrogram W contains 80 frequency bins and 30 source frames. The target spectrogram V contains the same frequency bins and 50 target frames. The mosaic is modeled as WH. How many scalar entries does the learned activation matrix H contain? **1500**

Question 3.15 NMF audio mosaicing:
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: the source spectrogram is fixed
Label **2**: constraints can encourage diagonal, smooth activations
Label **4**: the target spectrogram is fixed
Label **8**: activations are learned to represent the target using source frames **15**

Unit 4: Differentiable Digital Signal Processing (DDSP) /19 Points

Question 4.1 A scalar-gain model predicts $\hat{y} = gx$ for each component of an input vector. For input $(2, 3)$ and target $(6, 9)$, what value of g gives zero squared error? _____ **3**

Question 4.2 An FIR filter has coefficients $h(0) = 2$ and $h(1) = 1$. Its input samples are $x(0) = 1$ and $x(1) = 2$. Using $y(t) = \sum_k h(k)x(t-k)$, what is $y(1)$? _____ **5**

Question 4.3 An FIR filter has coefficients $h(0) = 1$, $h(1) = 2$, and $h(2) = 3$. Its input is $x(0) = 4$, $x(1) = 5$, and $x(2) = 6$. Using $y(t) = \sum_k h(k)x(t-k)$, what is the value of the final nonzero output sample? _____ **18**

Question 4.4 A 64-tap FIR filter has symmetric coefficients, $h(n) = h(63-n)$. How many independent coefficients can be trained? _____ **32**

Question 4.5 A model predicts samples $2, 1, 2, 0$ for a target signal $1, 3, 2, 2$. What is the mean squared error across the four samples? _____ **2.25**

Question 4.6 For one sample, a gain model predicts $\hat{y} = gx$ and uses squared loss $(\hat{y} - y)^2$. At $g = 2$, the input is $x = 3$ and the target is $y = 9$. What is the derivative of the loss with respect to g? _____ **-18**

Question 4.7 A gain model uses gradient descent with learning rate 0.1. Its current gain is $g = 2$, and the loss gradient is -12.
What is the updated gain after one step? _____ **3.2**

Question 4.8 Which statements describe DDSP?
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: DDSP combines DSP components with machine learning
Label **2**: DSP components are used only after neural-network training is complete
Label **4**: gradients can pass through a differentiable filter
Label **8**: a neural network can predict parameters for a DSP component _____ **13**

Question 4.9 A harmonic synthesizer uses $F_0 = 700$ Hz and sampling rate $F_s = 8000$ Hz. What is the largest harmonic number at or below the Nyquist frequency? _____ **5**

Question 4.10 Consider the control signals and components in a DDSP synthesizer.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: fundamental frequency controls the synthesized pitch
Label **2**: loudness can be provided as a synthesis control signal
Label **4**: harmonic weights control the room acoustics directly
Label **8**: a noise model can represent non-tonal sound components _____ **11**

Question 4.11 Waveform versus spectral-magnitude MSE:
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: waveform MSE compares audio samples
Label **2**: waveform MSE is phase-sensitive
Label **4**: spectral-magnitude MSE is less sensitive to small time shifts
Label **8**: spectral-magnitude MSE compares waveform samples **7**

Question 4.12 For an FIR filter, the prediction at time $t = 1$ is $\hat{y}(1) = h(0)x(1) + h(1)x(0)$. At the current parameters, $\hat{y}(1) = 5$, the target is $y(1) = 3$, and $x(0) = 3$. Using loss $(\hat{y}(1) - y(1))^2$, what is the loss derivative with respect to $h(1)$? **12**

Question 4.13 An FIR coefficient is currently $h(1) = 0.5$. Its loss gradient is 4, and gradient descent uses learning rate 0.1. What is the updated coefficient? **0.1**

Question 4.14 Consider harmonic synthesis and its limitations.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: a harmonic synthesizer uses a fundamental frequency and harmonic weights
Label **2**: harmonic weights act like gains for individual oscillators
Label **4**: a single harmonic source has multiple independent F_0 values
Label **8**: noise and irregular transients may require additional components **11**

Question 4.15 A harmonic synthesizer has $f_0 = 100$ Hz and four harmonic weights $[1, 2, 3, 4]$. It produces $y(t) = \sum_{k=1}^{4} a_k \sin(2\pi k f_0 t)$. What is $y(t)$ at $t = \frac{1}{4f_0}$? **-2**

Question 4.16 A sinusoidal synthesizer predicts $\hat{y} = 2\sin(\phi)$ and uses the squared loss $\mathcal{L} = (\hat{y} - 1)^2$. What is $\frac{\partial \mathcal{L}}{\partial \phi}$ at $\phi = 0$? **-4**

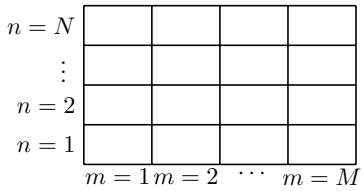
Question 4.17 A harmonic synthesizer predicts magnitudes $\hat{a}_1(g) = g$ at f_0 and $\hat{a}_2(g) = 2g$ at $2f_0$. The target magnitudes are $a_1 = 2$ and $a_2 = 3$. Using $\mathcal{L}_{\text{harm}}(g)) = \frac{1}{2}\left[(g-2)^2 + (2g-3)^2\right]$, what is $\frac{d\mathcal{L}_{\text{harm}}}{dg}$ at $g = 1$? **-3**

Question 4.18 Consider a noise synthesizer that filters white noise using a time-varying FIR filter.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: it can model breath noise and other stochastic sound components
Label **2**: changing the filter response can change the spectral shape and overall level of the noise
Label **4**: it provides precise, independent control over the phases of harmonic partials
Label **8**: the FIR filter length limits the spectral detail that can be modeled **11**

Question 4.19 Consider the design and interpretation of DDSP synthesizers.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: a musically motivated design can provide interpretable controls
Label **2**: a generic FIR filter is always the best model for learning room reverberation
Label **4**: DSP components must be chosen carefully to preserve interpretability and controllability
Label **8**: structured reverberation models can impose a stronger bias than a generic FIR filter
13

Unit 5: Differentiable Alignment Techniques for Music Processing /23 Points

In this unit, an $N \times M$ matrix, such as a cost matrix, is represented by a grid. The first index n increases upward, and the second index m increases to the right, as illustrated below. Thus, $(1, 1)$ is at the lower left and (N, M) is at the upper right.



Question 5.1 Let $C = \begin{array}{|c|c|c|c|}\hline 1&1&1&1\\\hline 1&1&1&1\\\hline\end{array}$. How many optimal warping paths exist? **3**

Question 5.2 Let $C = \begin{array}{|c|c|c|c|}\hline 1&1&1&1\\\hline 1&1&1&1\\\hline\end{array}$. What is the DTW distance? **4**

Question 5.3 Let $C = \begin{array}{|c|c|c|}\hline 2&1&2\\\hline 1&5&1\\\hline 1&1&4\\\hline\end{array}$. What is the DTW distance? **5**

Question 5.4 For $N = 2$ and $M = 3$,
how many valid alignment matrices exist? **5**

Question 5.5 For $N = 10$ and $M = 18$,
what is the minimum possible warping-path length? **18**

Question 5.6 For $N = 10$ and $M = 18$,
what is the maximum possible warping-path length? **27**

Question 5.7 For weakly aligned audio and target sequences:
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: framewise cross-entropy needs no frame labels
Label **2**: weak targets specify each frame's label
Label **4**: alignment loss handles alignment during training
Label **8**: alignment loss requires a known hard alignment **4**

Question 5.8 For the cost values $(4, 2, 4, 6)$, what value does the soft minimum with temperature parameter $\gamma > 0$ approach as $\gamma \to 0$? **2**

Question 5.9 For the cost values $(4, 2, 4, 6)$, what does the gradient weight assigned to the second cost approach as $\gamma \to 0$? **1**

Question 5.10 For the cost values $(1, 2, 3, 4, 5)$, what value does the gradient weight assigned to the second cost approach as $\gamma \to \infty$? **0.2**

Question 5.11 For the cost values $(4, 2, 6, 2)$, what does the gradient weight assigned to the second cost approach as $\gamma \to 0$? **0.5**

Question 5.12 Let $C = \begin{array}{|c|c|c|}\hline 4 & 1 & 2 \\ \hline 1 & 3 & 5 \\ \hline \end{array}$.

What is the SDTW value approached as $\gamma \to 0$? **4**

Question 5.13 Let $C = \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 0 & 0 \\ \hline \end{array}$. What is the expected alignment entry $E^{\gamma}(1,2)$?

Give a fraction as your answer. **1/3**

Question 5.14 Let $C = \begin{array}{|c|c|c|c|}\hline 1 & 2 & 3 & 4 \\ \hline \end{array}$.

For any $\gamma = 0.5$, what is the SDTW value? **10**

Question 5.15 For $\gamma = 0.1$, the SDTW value for an $N \times M$ cost matrix C is 2. Both endpoint costs, $C(1,1)$ and $C(N,M)$, are increased by 3, where $N, M > 1$. What is the new SDTW value? **8**

Question 5.16 A model parameter θ affects only one local cost, whose derivative with respect to θ is 2. The corresponding expected alignment entry is 0.2. What is the derivative of the SDTW loss with respect to θ? **0.4**

Question 5.17 Let $f(x,y) = 3x^2y + 2y^2$.

What is the value of $\frac{\partial f}{\partial x}$ at $(x,y) = (2,1)$? **12**

Question 5.18 Let $f(x,y) = (xy + 2x)^2$.

What is the value of $\frac{\partial f}{\partial x}$ at $(x,y) = (1,2)$? **32**

Question 5.19 Let $f(x,y) = x^2y^3$.

What is the value of $\frac{\partial^2 f}{\partial y\, \partial x}$ at $(x,y) = (2,1)$? **12**

Question 5.20 Let $f(x,y) = \min\{x^2 + y,\ x + 2y^2\}$.

What is the value of $\frac{\partial f}{\partial x}$ at $(x,y) = (2,1)$? **1**

Question 5.21 For SDTW on sequences of lengths N and M, consider runtime as both lengths grow.

Add the labels of all true statements (none: **0**; all: **15**).

Label **1**: the forward pass runs in $O(N + M)$ time

Label **2**: gradient computation runs in $O(N \log(N + M))$ time

Label **4**: the forward pass runs in $O(\log(NM))$ time

Label **8**: gradient computation runs in $O(N + M)$ time **0**

Question 5.22 Use $\text{softmin}(\mathbf{a}) = -\log \sum_i e^{-a_i}$.

Add the labels of all true statements (none: **0**; all: **15**).

Label **1**: $\text{softmin}(1,1) = 1$

Label **2**: $\text{softmin}(1,1) = 1 - \log 2$

Label **4**: $\text{softmin}(1,1) = \log 2$

Label **8**: $\text{softmin}(1,1) = 2$ **2**

Question 5.23 Use $\text{softmin}(\mathbf{a}) = -\log \sum_i e^{-a_i}$.
Add the labels of all true statements (none: **0**; all: **15**).
Label **1**: the softmin of positive inputs is always positive
Label **2**: the softmin is differentiable with respect to a_i
Label **4**: the softmin is at most the minimum input value
Label **8**: for $\mathbf{a} = (a_1)$, $\text{softmin}(\mathbf{a}) = a_1$

14