Some potentially useful formulas

$$
\begin{aligned}
F(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} \mathrm{d}t; \quad \omega \in \mathbb{R} \\
F_r &= \sum_{k=0}^{N_0-1} f_k e^{-irk\frac{2\pi}{N_0}}; \quad r \in \{0, \ldots, N_0 - 1\} \\
e^{ix} &= \mathrm{cox}(x) + i\mathrm{sin}(x) \\
L_p(\vec{x}, \vec{y}) &= \sqrt[p]{\sum_{d=1}^{D} |x_d - y_d|^p}
\end{aligned}
$$

1 The holistic approach to speech recognition

(10 pts)

You are given the following equations:

$$
\begin{aligned}
F_r &= \sum_{k=0}^{N_0-1} f_k \cos\left[\frac{\pi}{N_0}\left(k + \frac{1}{2}\right) r\right]; \quad r \in \{0, \ldots, N_0 - 1\} &(1) \\
\hat{w}_1^N &= \operatorname*{arg\,max}_{w_1^N} p(w_1^N | x_1^T) &(2) \\
p(x_1^T | w_1^N) &= \sum_{s_1^T} p(x_1^T, s_1^T | w_1^N) &(3) \\
f'_k &= f_k \cdot \left(0.54 - 0.46 \cos\left(\frac{2\pi k}{N - 1}\right)\right) &(4) \\
p(w_1^M) &= p(w_1) \prod_{m=2}^{M} p(w_m | w_{m-1}) &(5)
\end{aligned}
$$

To which module of the holistic approach does each of these equations belong:

- speech analysis,
- acoustic model,
- language model,
- search?

Justify your answer.

Hint: Each equation corresponds to exactly one module but not necessarily vice versa.

2 Word error rate (10 pts)

a) Calculate the word error rate when I speak into the iPhone

`oh my my oh my my`

and Siri recognizes

`my oh my my my`.

b) What is the (i) lowest and what the (ii) highest word error rate a speech recognizer can produce? How do spoken sentences and recognition hypotheses have to look like for (i) and (ii), respectively?

3 Fourier transform (16 pts)

a) Determine $F(0)$ for the following time signals

$$f_1(t) = \begin{cases} \sin(2\pi f_0 t) & : \quad 0 < t < \frac{1}{f_0} \\ 0 & : \quad \text{otherwise} \end{cases}$$

$$f_2(t) = \begin{cases} \sin(2\pi f_0 t) & : \quad 0 < t < \frac{1}{2f_0} \\ 0 & : \quad \text{otherwise} \end{cases}$$

$$f_3(t) = \begin{cases} 1 & : \quad 0 < t < \frac{1}{f_0} \\ 0 & : \quad \text{otherwise} \end{cases}$$

b) We have learned that sudden jumps in the time signal correspond to a wide range of frequencies in the spectrum. Suppose you have a discrete time signal

$$f_k = \begin{cases} 0 & : \quad k \in \{1, \ldots, N_0 - 1\} \\ 1 & : \quad k = 0. \end{cases}$$

Show that this discrete Dirac delta function has an entirely flat spectrum.

4 Speech analysis (6 pts)

The first two letters of the famous MFCC features stand for *Mel Frequency*, a scale $\tilde{f}$ that depends on the Hertz frequency scale f as

$$\tilde{f} = c \log_{10}\left(1 + \frac{f}{d}\right).$$

Determine the constants c and d such that

- 1 kMel corresponds to 1 kHz and
- c corresponds to 3 kHz.

5 Dynamic time warping (18 pts)

After speech analysis, an utterance spoken by a student results in the following feature vector sequence:

$$x_1^4 = \begin{bmatrix} 3 & 9 & 1 & 0 \\ 5 & 9 & 0 & 4 \end{bmatrix}.$$

The recognizer is able to recognize two words (`foo` and `bar`) each of which has a sample recording in the database:

$$\begin{aligned} x_1^{3,(\texttt{foo})} &= \begin{bmatrix} 8 & 3 & 0 \\ 9 & 8 & 3 \end{bmatrix}, \\ x_1^{5,(\texttt{bar})} &= \begin{bmatrix} 0 & 6 & 0 & 8 & 5 \\ 9 & 4 & 2 & 3 & 9 \end{bmatrix}. \end{aligned}$$

Determine the dynamic time warping costs $d'_{\texttt{foo}}$ and $d'_{\texttt{bar}}$ using the standard 0,1,2-alignment model and the Chebyshev distance L_∞. What did the student (presumably) say?