Midterm Exam

601.467/667 Introduction to Human Language Technology

Fall 2019
Johns Hopkins University
Co-ordinator: Philipp Koehn

9 October 2019

Q1. Auditory System and Speech Basics *15 points*

1. What are the approximate estimates of bit rates of speech signal and of linguistic message in speech and how would you estimate them?

2. How does human ear separate acoustic signal into its frequency components?

3. How is spectral resolution of human hearing dependent on frequency?

4. What is auditory masking and how can you observe it?

5. How do human speakers generate speech sounds?

Q2. Fourier Transform *10 points*

Recall that given a discrete-time signal $x[n]$, sampled at a rate of F_s, the short-time Fourier transform of the signal at time m is a complex-valued function

$$X[m](\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n},$$

where $w[\cdot]$ is the so called "window" function, $m, n \in \{\ldots, -1, 0, 1, \ldots\}$, and $-\pi < \omega \leq \pi$.

1. What sampling rates F_s are typically used in practice? Provide two standard values.

2. What window functions are typically used in practice? Name two standard choices, hand-draw a graph of $w[n]$ versus n for each, and specify the typical window sizes in milliseconds.

3. State in one sentence each the pros and cons of the two windows you have named. Hint: you may want to compare the sharpness of the main lobe and the "spectral smearing" due to the side lobes in the Fourier transform $W(\omega)$ of $w[n]$.

Q3. Spectrogram *10 points*

Recall that the spectrogram is a pictorial representation of $X[m](\omega)$, with m along the x-axis, ω along the y-axis, and the magnitude spectrum $|X[m](\omega)|$ plotted as the pixel intensity, darker for higher values.

1. What typical window sizes are associated with a *narrow-band* versus a *wide-band* spectrogram? Specify which window is wider, provide typical window sizes (in milliseconds), and relate the window size to the typical pitch period of human voice.

2. Name a tell-tale visual feature of a narrow-band spectrogram, and state what it represents. Hint: think horizontal striations.

3. Name a tell-tale visual feature of a wide-band spectrogram, and state what it represents. Hint: think vertical striations.

4. State what a *formant* is, and how it is visually manifested in a spectrogram.

Q4. Mel-Frequency Cepstral Coefficients *8 points*

Recall that a typical signal representation used for automatic speech recognition is a sequence of Mel-Frequency Cepstral Coefficients or MFCCs. In what order are the following signal processing operations carried out to compute MFCCs?

- ☐ Discrete cosine transform
- ☐ Discrete Fourier transform
- ☐ Logarithm computation
- ☐ Magnitude computation
- ☐ Mel-weighting (triangular filters of increasing width)
- ☐ Preemphasis
- ☐ Sampling the continuous-time signal
- ☐ Windowing

Finally, state one advantage of MFCCs over (linear) spectral magnitude for automatic speech recognition.

Q5. Classical Speech Recognition *10 points*

Recall that the classical formulation of the automatic speech recognition problem, inspired by information theoretic thinking, is the so called "source-channel" model, with channel input $\mathbf{W}$, channel output $\mathbf{A}$ and channel-decoder output $\widehat{\mathbf{W}}$.

1. Draw a block diagram to illustrate the source-channel model, and identify how human speech production and perception processes map on to your diagram:

   - the speaker's mind,
   - the speaker's vocal apparatus,
   - the transmission medium,
   - the listener's auditory apparatus, and
   - the listener's linguistic/cognitive facilities.

2. Mark parts of your block diagram that correspond to the component models of an automatic speech recognition system:

   - acoustic feature extraction,
   - the acoustic model, and
   - the language model.

   Where does the "search" function fall in this diagram?

3. Write down the Bayesian formulation of the speech recognition problem, sometimes referred to as the *fundamental equation* of speech recognition.

4. In the fundamental equation, identify the three core *computational* problems that must be solved to carry out automatic speech recognition: acoustic model estimation, language model estimation, and linguistic/hypothesis search.

5. Which of these three computational problems are solved offline, and which must be solved online, i.e. when carrying out recognition? What computational considerations/constraints does this imply for their solutions?

Q6. Hidden Markov Models *10 points*

Recall that hidden Markov models (HMMs) are widely used for acoustic modeling, and a major reason for their popularity is their compositional nature: HMMs for phonemes can be strung together to create HMMs for words, which in turn can be strung together to create HMMs for sentences, etc.

Using $\mathbf{x}$ to denote a sequence of observations $x_1, \ldots, x_T$ from an HMM, $\mathbf{s}$ the sequence $s_1, \ldots, s_T$ of unobserved states, and s_0 the (known) initial state, one may write

$$P(\mathbf{x}, \mathbf{s} \mid s_0) \;=\; \prod_{n=1}^{T} P(x_n|s_n)\, P(s_n|s_{n-1}).$$

State, both in words and in a mathematical precise expression, the computational problem associated with HMMs that is solved by

1. the forward-backward algorithm,

2. the Baum-Welch algorithm, and

3. the Viterbi algorithm.

For each algorithm above, specify its computational (big-O) complexity in terms of the input length T, and the number of states S in the HMM.

- If possible, refine your answer in terms of the number of permitted transitions (edges) E in the HMM instead of the number of states.

- Recall that in general $E = O(S^2)$. How are E and S related in typical HMMs used for acoustic modeling? Is this difference operationally significant?

Finally, state in words and via a simple formula, the mathematical approximation used to compute the n-gram language model probability

$$P(\mathbf{w}) \;=\; ?$$

of a word sequence $\mathbf{w} = w_1, w_2, \ldots, w_N$. What typical values of n are used in practice?

Q7. Speaker Recognition *7 points*

Explain the reasons why it is better to use Gaussian mixture model (multiple Gaussians) instead of a single Gaussian to model acoustic features (MFCCs) for speaker recognition?

Q8. Sigmoid Function *10 points*

The sigmoid function is defined as follows:

$$\sigma \triangleq \frac{1}{1+e^{-x}}$$

1. What is $\sigma(0)$, $\lim_{x\to\infty}\sigma(x)$, and $\lim_{x\to-\infty}\sigma(x)$?

2. Derive $\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x)(1-\sigma(x))$

3. Show that $\sigma'(x) \geq 0$

4. Plot the sigmoid function by hand by considering the above answers.

Q9. Softmax Function *10 points*

Suppose we have a J dimensional vector $\mathbf{h} \in \mathbb{R}^J$, the softmax function for element j is defined as follows:

$$[\text{softmax}(\mathbf{h})]_j \triangleq \frac{e^{h_j}}{\sum_{i=1}^{J} e^{h_i}}$$

1. Why the softmax function is used as a probabilistic distribution function? Please discuss it by considering $\sum_{j=1}^{J}[\text{softmax}(\mathbf{h})]_j$ and the range of $[\text{softmax}(\mathbf{h})]_j$.

2. Prove that the softmax function becomes the sigmoid function when $J = 2$ by setting $h_2 - h_1 = -x$ (or $h_1 - h_2 = -x$)

Q10. Chain Rule and Back Propagation *10 points*

The chain rule is defined as follows:

$$\frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$$

1. Derive the derivative of $\log(\sigma(x))$. Here $\log'(x) = \frac{1}{x}$.

2. Derive the partial derivative of $\sigma(ax + b)$ with respect to a and b

3. One of the most powerful techniques of the deep neural network is that the model parameters are efficiently estimated by a back propagation algorithm. State a couple of sentences about the back propagation algorithm based on the chain rule.