ECE 537 Fundamentals of Speech Processing
Problem Set 7

UNIVERSITY OF ILLINOIS
Department of Electrical and Computer Engineering

Assigned: Sunday, 10/23/2022; Due: Friday, 10/28/2022
Reading: Narendranath, Murthy and Yegnanarayan, "Transformation of formants for voice conversion using artificial neural networks," 1995

1. (a) (1 point) Recall that LPC finds coefficients a_k such that $v[n]$ is approximately white (specifically, so that $v[n] \perp s[n-k]$), where

$$v[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k]$$

The air pressure just above the glottis, however, is not quite quite noise; it more closely resembles an LF model. Let us use $p[n] = U_g'(tF_s)$ to denote the samples of the LF model, $U_g'(t)$, at a sampling rate of F_s samples/second. The DTFT of the LF model is complicated, but at most frequencies, $|P(\omega)|$ is well modeled as the frequency response of an all-pole filter with two poles: one at zero frequency with a bandwidth of α radians/second, and one at zero frequency with a bandwidth of ϵ radians/second, where α and ϵ are standard parameters of the LF model denoting the rate of glottal opening and the rate of glottal closing, respectively. Suppose that you wish to model $p[n]$ by the signal

$$p[n] = v[n] + a_1 p[n-1] + a_2 p[n-2],$$

where $v[n]$ is an impulse train. For what values of a_1 and a_2 will this model have the correct magnitude frequency response, $|P(\omega)|$? Express your answer in terms of α, ϵ, and/or F_s.

(b) (1 point) Recall that, during production of the vowel /ə/, the vocal tract is roughly a uniform tube closed at one end and open at the other, and therefore the formant frequencies are

$$F_k = \frac{c}{4L} + (k-1)\frac{c}{2L},$$

where L is the length of the vocal tract, and $c = 354$m/s is the speed of sound at body temperature. For other vowels, the formants vary, but with few exceptions (e.g., F2 in /i,u,o/ and F3 in /r/), the formant frequencies stay within the following bounds:

$$(k-1)\frac{c}{2L} \leq F_k \leq k\frac{c}{2L}$$

Suppose that you want to model the speech signal using a 10^{th}-order LPC synthesis filter of the form

$$\hat{s}[n] = v[n] + \sum_{m=1}^{10} a_m \hat{s}[n-m],$$

where $v[n]$ is an impulse train. Find the sampling rate F_s, as a function of L, for which this model is most accurate.

2. This problem explores exactly the sense in which cross-entropy is better than mean-squared-error for classification problems. Suppose that we have a training corpus with only one training token. For this one training token, the correct label is $\vec{\zeta} = [1, 0, \ldots, 0]^T$, i.e., $\zeta_1 = 1$, and $\zeta_k = 0$ for $k \neq 1$. The neural network computes some logits y_j, then computes the softmax output according to

$$z_i = \frac{e^{y_i}}{\sum_{j=1}^{N} e^{y_j}}$$

Notice that z_i depends on y_j for all k, not just for $k = i$. Therefore, to compute the gradient of the error ($\mathcal{E}$) with respect to y_k, you need to compute

$$\frac{\partial \mathcal{E}}{\partial y_k} = \sum_{i=1}^{N} \left(\frac{\partial \mathcal{E}}{\partial z_i} \right) \left(\frac{\partial z_i}{\partial y_k} \right)$$

(a) (1 point) Suppose that the error is MSE, thus

$$\mathcal{E} = \frac{1}{2} \sum_{i=1}^{N} (z_i - \zeta_i)^2$$

Express $\frac{\partial \mathcal{E}}{\partial y_k}$ as a function of the softmax outputs z_i, for $1 \leq i \leq N$ and $1 \leq k \leq N$.

(b) (1 point) Now suppose that the error is cross-entropy. Since the correct answer is $\zeta_1 = 1$, the cross-entropy loss for this training dataset is just

$$\mathcal{E} = -\ln z_1$$

Express $\frac{\partial \mathcal{E}}{\partial y_k}$ as a function of the softmax outputs z_i, for $1 \leq i \leq N$ and $1 \leq k \leq N$.