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.

Solution: We want a filter with two poles, both at zero frequency, with bandwidths of α and ϵ radians/second, respectively. That filter is

$$G(\omega) = \frac{1}{(1 - e^{-\alpha/F_s} z^{-1})(1 - e^{-\epsilon/F_s} z^{-1})}$$

Setting this equal to the frequency response of a second-order LPC synthesis filter, $\frac{1}{1-a_1 z^{-1} - a_2 z^{-2}}$, we find that

$$a_1 = e^{-\alpha/F_s} + e^{-\epsilon/F_s}$$
$$a_2 = -e^{-(\alpha+\epsilon)/F_s}$$

(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.

Solution: A 10^{th}-order LPC synthesis filter has 10 poles. Two of those poles are required to model the non-whiteness of the source spectrum (specifically, to model α and ϵ in the LF model). The other 8 poles can be used to model 4 formant frequencies (for each formant frequency, we need a complex-conjugate pole pair). Therefore, this model is most accurate if the Nyquist rate is $F_n = kc/2L$, thus the sampling rate is

$$F_s = 2F_N = 2 \times \left(k\frac{c}{2L}\right)$$

for values of $k = 4$ and $c = 354$, i.e.,

$$F_s = \frac{4c}{L}$$

If the sampling rate is larger, then there might be five formants to model, which is too many for our model. If the sampling rate is smaller, then there might be only three formants to model, which is too few for our model.

For example, some large men have vocal tracts (including both the mouth and the pharynx, from glottis to lips) as long as $L = 0.177$m, for which

$$c = \frac{4 \times 354}{0.177} = 8000\text{Hz}$$