Frequency-Domain Representations

Introduction

The speech production model shown in Figure 1 is a linear system for either voiced or unvoiced speech.

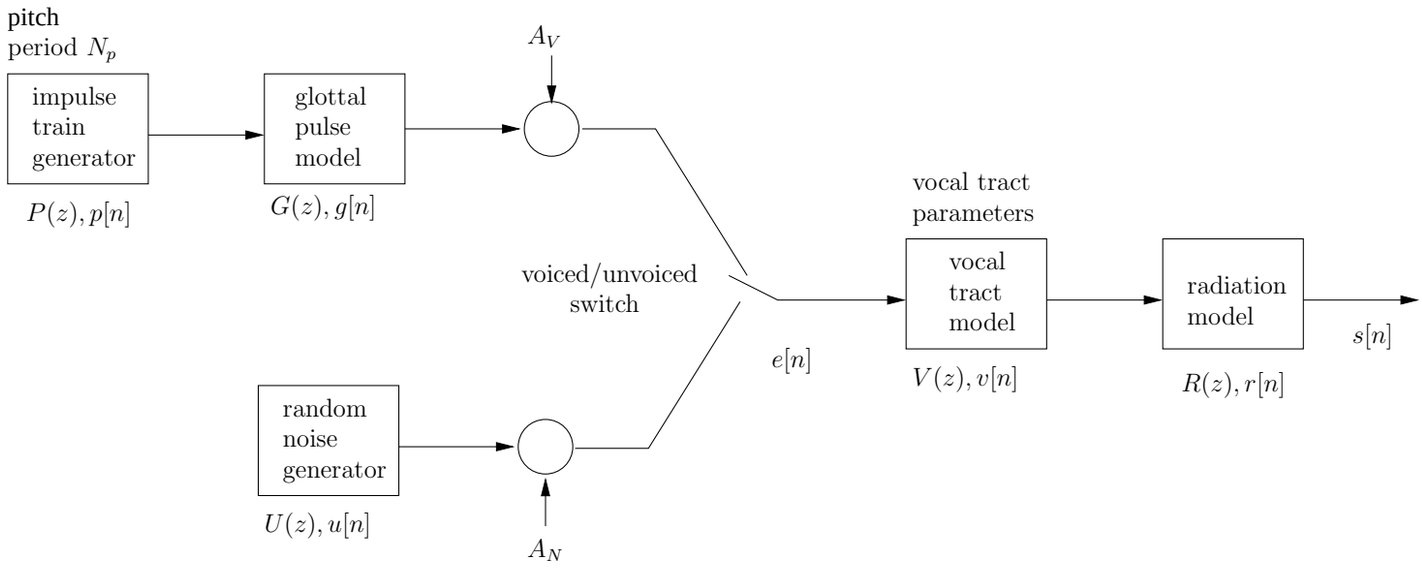

Figure: Speech Production Model

For a sustained voiced speech, we have

$$S(e^{j\omega}) = A_V P(e^{j\omega}) G(e^{j\omega}) V(e^{j\omega}) R(e^{j\omega}) \tag{1}$$

For a sustained unvoiced speech, assuming white noise with unit power, we have

$$\Phi_{ss}(e^{j\omega}) = A_N^2 |V(e^{j\omega})|^2 |R(e^{j\omega})|^2 \tag{2}$$

So it is clear that the discrete-time Fourier transform of the output signal is dependent on the input signal.

Short-time Fourier Transform

Since the temporal properties of a speech signal changes every $10 - 40$ msec, we should instead use the short-time Fourier transform (STFT).

Discrete-time Fourier Analysis

Discrete-time Fourier Transform and the Inverse

Recall that DTFT and IDTFT are defined as

$$X(e^{j\omega}) = \sum_n x[n] e^{-j\omega n}, \quad x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega n} d\omega. \tag{3}$$

Furthermore, the DFT and IDFT are defined as

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j(2\pi k/N)n}, \quad k = 0, 1, \ldots, N-1$$
$$x[n] = \frac{1}{N}\sum_{n=0}^{N-1} X[k]e^{j(2\pi k/N)n}, \quad n = 0, 1, \ldots, N-1 \tag{4}$$

For a finite-length sequence,

$$X[k] = X(e^{j\omega})|_{\omega=2\pi k/N} \tag{5}$$

An Important Example: Periodic Impulse Train

Consider the signal

$$p[n] = \sum_r \delta[n - rN], \tag{6}$$

which is a periodic impulse train with period N. By (4), the DFT of $p[n]$ is 1. So we have (also by (4))

$$p[n] = \frac{1}{N}\sum_{k=0}^{N-1} e^{j\frac{2\pi k}{N}n}. \tag{7}$$

Thus we have

$$P(e^{j\omega}) = \frac{2\pi}{N}\sum_{k=0}^{N-1} \delta(\omega - \frac{2\pi k}{N}) \tag{8}$$

since the DTFT of a complex complex exponential signal $e^{j\omega_0 n}$ is $2\pi\delta(\omega - \omega_0)$.

Short-time Fourier Analysis

The short-time (time-dependent) Fourier transform is defined as

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_m w[\hat{n} - m]x[m]e^{-j\hat{\omega}m}, \tag{9}$$

which is the DTFT of a window of $x[n]$ around $\hat{n}$. Alternatively, by change of variable, (9) is equivalent to

$$\begin{aligned} X_{\hat{n}}(e^{j\hat{\omega}}) &= \sum_{m'} w[m']x[\hat{n} - m']e^{-j\hat{\omega}(\hat{n}-m')} \\ &= e^{-j\hat{\omega}\hat{n}} \sum_m x[\hat{n} - m]w[m]e^{j\hat{\omega}m}. \end{aligned} \tag{10}$$

We can express (10) as

$$X_{\hat{n}}(e^{j\hat{\omega}}) = e^{-j\hat{\omega}\hat{n}}\tilde{X}_{\hat{n}}(e^{j\hat{\omega}}), \tag{11}$$

where

$$\begin{aligned} \tilde{X}_{\hat{n}}(e^{j\hat{\omega}}) &\triangleq \sum_m x[\hat{n} - m]w[m]e^{j\hat{\omega}m} \\ &= \sum_m x[\hat{n} + m]w[-m]e^{-j\hat{\omega}m}. \end{aligned} \tag{12}$$

Note that $\tilde{X}_{\hat{n}}(e^{j\hat{\omega}})$ can be seen as a function of $\hat{\omega}$ for fixed $\hat{n}$ (the DTFT perspective), or a function of $\hat{n}$ for fixed $\hat{\omega}$ (the linear filtering perspective).

DTFT Interpretation

By (9), $X_{\hat{n}}(e^{j\hat{\omega}})$ is the DTFT of the sequence $w[\hat{n}-m]x[m]$. Applying the inverse DTFT, we have

$$w[\hat{n}-m]x[m] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X_{\hat{n}}(e^{j\hat{\omega}})e^{j\hat{\omega}m}d\hat{\omega}. \tag{13}$$

It follows that

$$x[\hat{n}] = \frac{1}{2\pi w[0]}\int_{-\pi}^{\pi} X_{\hat{n}}(e^{j\hat{\omega}})e^{j\hat{\omega}n}d\hat{\omega}. \tag{14}$$

(14) is an important theoretical result, showing how to reproduce $x[n]$ from STFT in principle. So $x[n]$ and $\tilde{X}_{\hat{n}}(e^{j\hat{\omega}})$ are equivalent. However, computationally it is not very useful since it involves an integral. It can be shown that a discrete version of STFT, short-time discrete Fourier transform $\tilde{X}_{\hat{n}}[k]$ can be used to represent $x[n]$.

Since time-domain multiplication leads to frequency-domain convolution, we have

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{-j\omega})e^{-j\omega\hat{n}}X(e^{j(\hat{\omega}-\omega)})d\omega, \tag{15}$$

where $W(e^{-j\omega})e^{-j\omega\hat{n}}$ is the DTFT of $w[\hat{n}-m]$. Clearly, the spectrum $W(e^{j\omega})$ is influential on the short-time spectrum. In terms of spectral resolution, the longer the window, the better the resolution since the main lobe width of $W(e^{j\omega})$ is proportional to $\frac{1}{L}$.

Linear Filtering Interpretation

(9) can be re-written as

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_m w[\hat{n}-m]\left(x[m]e^{-j\hat{\omega}m}\right). \tag{16}$$

We can view $X_{\hat{n}}(e^{j\hat{\omega}})$ in (16) for fixed $\hat{\omega}$ as a time sequence of complex numbers indexed by $\hat{n}$. Specifically, it is obtained by the convolution (i.e. filtering) of the sequences $w[n]$ and $x[n]e^{-j\hat{\omega}n}$.
Yet another perspective of linear filtering is (10).

Spectrographic Displays

Spectrogram

An image for STFT. Each slice represents the spectrum of an analysis frame.

Wideband Spectrogram

With a short analysis window, the spectral resolution is broad and the spectrum is smooth. The variation from window to window is clear, leading to vertical striations.

Narrowband Spectrogram

With a long analysis window, the spectral resolution is detailed and the spectrum is not smooth.

Let R be the sampling period in time of STFT, and L be the length of analysis window. We have, from (9),

$$X_{rR}(e^{j\omega}) = \sum_{m=rR-L+1}^{rR} w[rR-m]x[m]e^{-j\omega m}, \tag{17}$$

Suppose we have N sampling points

$$X_{rR}(e^{j\omega_k}), \quad \omega_k = 2\pi k/N, \quad k = 0, 1, \ldots, N-1$$

in the frequency domain. $X[k] \triangleq X_{rR}(e^{j\omega_k})$ are the DFT of $y[m] \triangleq w[rR-m]x[m]$, and therefore

$$y[m] = w[rR-m]x[m] = \frac{1}{N}\sum_{k=0}^{N-1} X_{rR}[k]e^{j\omega_k m}. \tag{18}$$

So in principle, $x[m]$ can be reconstructed from STFT, as long as

$$R \leq L \leq N \tag{19}$$

to ensure every point is included in at least a window, and the number of frequency-domain sampling points is sufficient.

OLA Synthesis

The basic equation is

$$y[n] = \sum_r \left(\frac{1}{N}\sum_{k=0}^{N-1} Y_r[k]e^{j\omega_k n} \right), \tag{20}$$

where $Y_r[k] \triangleq X_{rR}(e^{j\omega_k})$. Defining the reconstruction signal of the rth analysis window

$$y_r[n] = \frac{1}{N}\sum_{k=0}^{N-1} Y_r[k]e^{j\omega_k n} = x[n]w[rR-n] \tag{21}$$

we have

$$y[n] = \sum_r y_r[n] = x[n]\sum_r w[rR-n] \triangleq x[n]\tilde{w}[n]. \tag{22}$$

For exact construction, we need that

$$\tilde{w}[n] = \sum_r w[rR-n] = C, \tag{23}$$

where C is called the reconstruction gain.