Last lecture I introduced the idea that any function defined on \( x \in 0, \ldots, N-1 \) could be written a sum of sines and cosines. There are two different reasons why this is useful. The first is a general one, that sines and cosines behave nicely under convolution and so we can sometimes understand better what filtering does if we understand its effects on sines and cosines. The second is more specific, that sines and cosines are a natural set of functions for describing sounds.

Today I will begin with the basic theory of Fourier analysis. This is a particular way of writing a signal as a sum of sines and cosines.

**Discrete Fourier Transform**

Consider 1D signals \( I(x) \) which are defined on \( x \in \{0,1,\ldots,N-1\} \). Define the \( N \times N \) Fourier transform matrix \( \mathbf{F} \) whose \( k^{th} \) row and \( x^{th} \) column is:

\[
\mathbf{F}_{k,x} = \cos\left(\frac{2\pi}{N} k x\right) - i \sin\left(\frac{2\pi}{N} k x\right) = e^{-i \frac{2\pi}{N} k x}
\]

Note that this matrix is symmetric since \( e^{-i \frac{2\pi}{N} k x} = e^{-i \frac{2\pi}{N} x k} \). Also note that each row and column of the matrix \( \mathbf{F} \) has a real part and an imaginary part. The real part is a sampled cosine function. The imaginary part is a sampled sine function. Note that the leftmost and rightmost column of the matrix \( x = 0 \) and \( x = N-1 \) are not identical. You would need to go to \( x = N \) to reach the same value as at \( x = 0 \), but \( x = N \) is not represented. Similarly, the first and last row \( k = 0 \) and \( k = N-1 \) are not identical.

Right multiplying the matrix \( \mathbf{F} \) by the \( N \times 1 \) vector \( I(x) \) gives a vector \( \hat{I}(k) \)

\[
\hat{I}(k) \equiv \mathbf{F} I(x) = \sum_{x=0}^{N-1} I(x)e^{-i \frac{2\pi}{N} k x}
\]

which is called the discrete Fourier transform of \( I(x) \). In general, \( \hat{I}(k) \) is a complex number for each \( k \). We can write it using Euler’s equation:

\[
\hat{I}(k) = A(k)e^{i\phi(k)}
\]

\( |\hat{I}(k)| = A(k) \) is called the amplitude spectrum and \( \phi(k) \) is called the phase spectrum.

**Inverse Fourier transform**

One can show (see Appendix A) that

\[
\mathbf{F}^{-1} = \frac{1}{N} \mathbf{F}
\]

where \( \mathbf{F} \) is the matrix of complex conjugates of \( \mathbf{F} \).

\[
\mathbf{F}_{k,x} \equiv e^{i \frac{2\pi}{N} k x}.
\]

So, \( \frac{1}{N} \mathbf{F} \mathbf{F} \) is the identity matrix.
Periodicity properties of the Fourier transform

The Fourier transform definition assumed that the function was defined on $x \in 0, \ldots, N - 1$, and for frequencies $k$ in $0, \ldots, N - 1$. However, sometimes we will want to be more flexible with our range of $x$ and $k$.

For example, we may want to consider functions $h(x)$ that are defined on negative values of $x$ such as the local difference function $D(x)$, the local average function $B(x)$, the Gaussian function which has mean 0, Gabor functions, etc. The point of the Fourier transform is to be able to write a function as a sum of sinusoids. Since sine and cosine functions are defined over all integers, there is no reason why the Fourier transform needs to be defined only on functions that are defined on $x$ in 0 to $N - 1$.

We can define the Fourier transform of any function that is defined on a range of $N$ consecutive values of $x$. For example, if we have a function defined on $-\frac{N}{2}, \ldots, -1, 0, 1, \frac{N}{2} - 1$, then we can just write the Fourier transform as

$$
\hat{I}(k) \equiv \mathbf{F} I(x) = \sum_{x=-\frac{N}{2}}^{\frac{N}{2}-1} h(x) e^{-i \frac{2\pi}{N} kx}
$$

Essentially what we are doing here is treating this function $h(x)$ as periodic with period $N$, just like sine and cosine are, and compute the Fourier transform over a convenient sequence of $N$ sample points. Later this lecture I will calculate the Fourier transform of $D(x)$ and $B(x)$, so look ahead to see how that is done.

The second aspect of periodicity in the Fourier transform is that $\hat{I}(k)$ is well-defined for any integer $k$ (cycles per $N$ pixels). The definition of the Fourier transform doesn’t just allow $k$ in 0 to $N - 1$, but rather $k$ can be any integer. In that case, $\hat{I}(k)$ may be considered periodic in $k$ with period $N$,

$$
\hat{I}(k) = \hat{I}(k + mN)
$$

since, for any integer $m$,

$$
e^{i 2\pi m} = \cos(2\pi m) + i \sin(2\pi m) = 1
$$

and so

$$
e^{i \frac{2\pi}{N} kx} = e^{i \frac{2\pi}{N} k} e^{i \frac{2\pi}{N} mN} = e^{i \frac{2\pi}{N} (k + mN)x}
$$

Thus, if we use frequency $k + mN$ instead of $k$ in the definition of the Fourier transform, we get the same value.

Conjugacy property of the Fourier transform

It is a bit strange that our function $I(x)$ has $N$ points and we will write it in terms of $2N$ functions, namely $N$ cosines and $N$ sines. I mentioned this point last lecture as well, and showed that indeed only $N$ functions are needed, namely $\frac{N}{2} + 1$ cosines and $\frac{N}{2} - 1$ sines. This suggests that there is a redundancy in $\hat{I}(k)$ values. The redundancy is that $\cos(\frac{2\pi}{N} kx) = \cos(\frac{2\pi}{N} (N - k)x)$ and so taking the inner product with $I(x)$ will give the same value for frequency $k$ as $N - k$. Similarly, $\sin(\frac{2\pi}{N} kx) = -\sin(\frac{2\pi}{N} (N - k)x)$ and so taking the inner product of $I(x)$ with these two functions will give the same value but with opposite sign.
Conjugacy property: If $I(x)$ is a real valued function, then

$$\hat{I}(k) = \hat{I}(N - k).$$

The property does not apply if $I(x)$ has imaginary components. We will see an example later, namely if we take the Fourier transform of $e^{i\frac{2\pi}{N}k_0x}$, for some fixed frequency $k_0$.

For the proof of the Conjugacy Property, see Appendix B.

Linear Filtering

The visual and auditory systems analyze signals by filtering them into bands (ranges of different frequencies) of sines and cosines. The idea of a filter should be intuitive to you. You can imagine having a large bag of rocks and wanting to sort the rocks into ranges of different sizes. You could first pass the rocks through a fine mesh that has small holes only, so only the small rocks would pass through. Then take the bigger rocks that didn’t pass through, and pass them through a mesh filter that has slightly larger holes so that now the medium size rocks pass through, but not the large rocks. This would give you three sets of rocks of a different range of sizes.

You are also intuitively familiar with filtering from color vision where the L, M, and S receptors selectively absorb the incoming light by wavelength. There is some frequency overlap in the sensitivity functions, so we don’t have a perfect separation of frequency bands by the three cones.

The figure below shows a more concrete example of the filtering that we will be considering. Here we have 1D signal in the upper left panel. We can write this signal as a sum of signals that have different ranges of frequencies. In this example, the original signal is exactly the sum of the other five signals. We will see shortly how this can be done.

Convolution Theorem

A very useful property of the Fourier transform is the Convolution Theorem: for any two functions $I(x)$ and $h(x)$ that are defined on 0 to $N - 1$,

$$\mathbf{F}(I(x) * h(x))) = \mathbf{F}I(x) \mathbf{F}h(x) = \hat{I}(k) \hat{h}(k).$$

For the proof see Appendix C.

To prove this theorem, we need to deal with a similar issue that we mentioned before that the functions might be defined on values of $x$ other than 0 to $N - 1$. We do so by assuming the functions are periodic i.e. $I(x) = I(x + mN)$ and $h(x) = h(x + mN)$ for any integer $m$ and we define the summation from 0 to $N - 1$.

Filtering and bandwidth

Suppose we convolve an image $I(x)$ with a function $h(x)$. We have referred to $h(x)$ as an impulse response function. $h(x)$ is also called a linear filter. Recall that the Fourier transform of the filter $h(x)$ can be written

$$\hat{h}(k) = |\hat{h}(k)| e^{i\phi(k)}$$

or frequency i.e. since light travels at a constant speed (called $c$), we can equivalently describe the sensitivity of L, M, and S cones to frequency (either spatial frequency $\lambda$ or temporal frequency $\omega$, where $c = \omega \lambda$.

---

1 last updated: 12th Apr, 2018

©Michael Langer
where $|\hat{h}(k)|$ is called the amplitude spectrum and $\phi(k)$ is called the phase spectrum. By the convolution theorem,

$$F I(x) = F(I(x) \ast h(x)) = \hat{I}(k) \, |\hat{h}(k)| \, e^{i\phi(k)}$$

and $|\hat{h}(k)|$ amplifies or attenuates the frequency component amplitude $|\hat{I}(k)|$ and the phase $\phi(k)$ of the filter shifts each frequency component.

We can characterize filters by how they affect different frequencies. We will concern ourselves mainly with the amplitude spectrum for now. Let’s first address the case of “ideal” filters. We say:

- $h(x)$ is an ideal low pass filter if there exists a frequency $k_0$ such that

$$\hat{h}(k) = \begin{cases} 1, & 0 \leq k \leq k_0 \\ 0, & k_0 < k \leq \frac{N}{2} \end{cases}$$

- $h(x)$ is an ideal high pass filter if there exists $k_0$ such that

$$\hat{h}(k) = \begin{cases} 0, & 0 \leq k < k_0 \\ 1, & k_0 \leq k \leq \frac{N}{2} \end{cases}$$

- $h(x)$ is an ideal bandpass filter if there exists two frequencies $k_0$ and $k_1$ such that

$$\hat{h}(k) = \begin{cases} 0, & 0 \leq k < k_0 \\ 1, & k_0 \leq k \leq k_1 \\ 0, & k_1 < k \leq \frac{N}{2} \end{cases}$$

Note that these definitions above only concern $k \in \{0, \ldots, \frac{N}{2}\}$. Frequencies $k < 0$ and frequencies $k > \frac{N}{2}$ are ignored in the definition because the values of $\hat{h}(k)$ of these frequencies are determined by the conjugacy and periodicity properties.
Non-ideal filters and bandwidth

We typically work with filters that are not ideal i.e. filters that only approximately satisfy the above definitions. If we have an approximately bandpass filter, then we would like to describe the width of this filter i.e. the range of frequencies that it lets through. One often does this by considering the frequencies at which $|\hat{h}(k)|$ reaches half its maximum value. The bandwidth at half-height is defined to be $k_1 - k_0$, where $k_0 < k_1$ and

$$|\hat{h}(k_0)| = |\hat{h}(k_1)| = \frac{1}{2} \max_{k \in [0, N/2]} |\hat{h}(k)|$$

Bandwidth can also be defined in terms of the ratio of $k_1$ to $k_0$, specifically, the octave bandwidth at half height is:

$$\log_2\left(\frac{k_1}{k_0}\right) = \log_2(k_1) - \log_2(k_0)$$

For example, a filter with a bandwidth of one octave means that the $k_1$ frequency is twice the $k_0$ frequency.

Examples of filters and their Fourier transforms

Let’s look at some examples, starting with an impulse function, and the local difference and local average. Some of our calculations of Fourier transforms below will use Euler’s formula, $e^{i\theta} = \cos(\theta) + i\sin(\theta)$. In particular, you can verify for yourselves that:

$$\cos(\theta) = \frac{1}{2}(e^{i\theta} + e^{-i\theta})$$

$$i\sin(\theta) = \frac{1}{2}(e^{i\theta} - e^{-i\theta})$$

We will often take $\theta = \frac{2\pi}{N} k x$.

Example 1: Impulse function

Recall

$$\delta(x) \equiv \begin{cases} 1, & x = 0 \\ 0, & \text{otherwise} \end{cases}$$
Its Fourier transform is

\[ \hat{\delta}(k) = \sum_{x=0}^{N-1} \delta(x) e^{-i \frac{2\pi}{N} k x} \]
\[ = 1 \cdot e^{i \frac{2\pi}{N} k 0} \]
\[ = 1 \]

This is rather surprising. It says that an impulse function can be written as sum of cosine functions over all frequencies \( k \in [0, 1, \ldots, N-1] \) and dividing by \( N \), i.e.

\[ \delta(x) = \frac{1}{N} \sum_{x=0}^{N-1} \hat{\delta}(k) e^{i \frac{2\pi}{N} k x} \]

Note that I write cosine functions, rather than cosine and sine functions, since \( \hat{\delta}(k) = 1 \) and so the phase is 0, i.e. \( \phi(k) = 0 \) for all \( k \), i.e. purely real, and so there are no sine (imaginary) components. Basically, what happens is that all the cosine functions have the value 1 at \( x = 0 \), whereas at other values of \( x \) there are a range of values, some positive and some negative, and these other values cancel each other out when you take the sum.

To try to illustrate what is going on here, I have written a Matlab script

http://www.cim.mcgill.ca/~langer/546/MATLAB/sumOfSinusoids.m

which shows what happens when you add up all the cosines (top) and sines (bottom) of frequency \( k = 0, \ldots, N-1 \) for some chosen \( N \).

**Example 2: local difference**

Recall the local difference function \( D(x) \) from last lecture. It has value \(-\frac{1}{2}\) at \( x = 1 \) and value \( \frac{1}{2} \) at \( x = -1 \). Let’s compute its Fourier transform.

\[ \hat{D}(k) = \sum_{x} D(x) e^{-i \frac{2\pi}{N} k x} \]
\[ = \frac{1}{2}(-1 \cdot e^{-i \frac{2\pi}{N} k} + 1 \cdot e^{-i \frac{2\pi}{N} k(-1)}) \]
\[ = \frac{1}{2}(-e^{-i \frac{2\pi}{N} k} + e^{i \frac{2\pi}{N} k}) \]
\[ = i \sin\left(\frac{2\pi}{N} k\right) \]

Notice that \( \hat{D}(k) \) is purely imaginary and the plot below shows the imaginary component only. The phase spectrum is constant \( e^{i \frac{\pi}{2}} = \frac{\pi}{2} \).
Example 3: local average

\[ B(x) = \begin{cases} 
\frac{1}{2}, & x = 0 \\
\frac{1}{4}, & x = -1 \\
\frac{1}{4}, & x = 1 \\
0, & \text{otherwise}
\end{cases} \]

Taking its Fourier transform,

\[
\mathbf{F} B(x) = \frac{1}{2} + \frac{1}{4} (e^{-i \frac{2\pi}{N} k} + e^{-i \frac{2\pi}{N} (-1)}) \\
= \frac{1}{2} + \frac{1}{4} (e^{-i \frac{2\pi}{N} k} + e^{i \frac{2\pi}{N} k}), \\
= \frac{1}{2} (1 + \cos(\frac{2\pi}{N} k))
\]

Notice that \( \hat{B}(k) \) is real, i.e. it has no imaginary component. Moreover it is non-negative. Thus, the phase spectrum \( \phi(k) \) is 0.

Example 4: the “complex exponential”

Let \( h(x) = e^{i \frac{2\pi}{N} k_0 x} \) for some integer frequency \( k_0 \). Then,

\[
\mathbf{F} e^{i \frac{2\pi}{N} k_0 x} = N\delta(k - k_0).
\]

See the Appendix A for a proof.

Is this result surprising. In hindsight, no. Taking the Fourier transform of a function amounts to finding out what are the coefficients on the complex exponentials \( e^{i \frac{2\pi}{N} k x} \) for various \( k \) such that you can add these complex exponentials up and get the function. But if the function itself is a single complex exponential, then there is just one non-zero complex exponential needed!

We will use this result below when we compute the Fourier transforms of a cosine and sine function.
Example 5: constant function \( h(x) = 1 \)

This is just a special case of the last example, namely if we take \( k_0 = 0 \). In this case,

\[
\hat{h}(k) = N \delta(k).
\]

Thus, the Fourier transform of the constant function \( h(x) = 1 \) is a delta function *in the frequency domain*, namely it has value \( N \) at \( k = 0 \) and has value 0 for all values of \( k \) in \( 1, \ldots, N - 1 \).

Examples 6 and 7: cosine and sine

We use Euler’s formula to rewrite cosine and sine as a sum of complex exponentials.

\[
\mathbf{F} \cos\left(\frac{2\pi}{N} k_0 x\right) = \sum_{x=0}^{N-1} \cos\left(\frac{2\pi}{N} k_0 x\right)e^{-i \left(\frac{2\pi}{N} k x\right)}
\]

\[
= \sum_{x=0}^{N-1} \frac{1}{2} (e^{i \frac{2\pi}{N} k_0 x} + e^{-i \frac{2\pi}{N} k_0 x})e^{-i \frac{2\pi}{N} k x}
\]

\[
= \frac{N}{2} (\delta(k_0 - k) + \delta(k_0 + k))
\]

\[
\mathbf{F} \sin\left(\frac{2\pi}{N} k_0 x\right) = \sum_{x=0}^{N-1} \sin\left(\frac{2\pi}{N} k_0 x\right)e^{-i \left(\frac{2\pi}{N} k x\right)}
\]

\[
= \sum_{x=0}^{N-1} \frac{1}{2i} (e^{i \frac{2\pi}{N} k_0 x} - e^{-i \frac{2\pi}{N} k_0 x})e^{-i \frac{2\pi}{N} k x}
\]

\[
= -\frac{N}{2i} (\delta(k_0 - k) - \delta(k_0 + k))
\]

Example 7: Gaussian

If we sample a Gaussian function

\[
G(x, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}}
\]

on integer values of \( x \), and take the Fourier transform, we get the following approximation:

\[
\hat{G}(k, \sigma) \approx e^{-\frac{1}{2} \left(\frac{k}{N}\right)^2 \sigma^2 k^2}
\]

This approximation becomes exact in the limit as \( N, \sigma \rightarrow \infty \), with \( \frac{\sigma}{N} \) held constant. (This amounts to taking the continuous instead of discrete Fourier transform. The proof of these claims are beyond the scope of this course.)

If you wish to see this approximation for yourself, run the Matlab script

http://www.cim.mcgill.ca/~langer/546/MATLAB/plotFourierTransformGaussian.m

which generates the figure
A few key properties to notice are:

- If the standard deviation of the Gaussian in the space \((x)\) domain is \(\sigma\) then the standard deviation of the Gaussian in the frequency \((k)\) domain is proportional to \(\frac{1}{\sigma}\).

- \(\hat{G}(k, \sigma)\) has a Gaussian shape, but it does not integrate to 1, namely there is no scaling factor present. The max value occurs at \(k = 0\) and the max value is always 1.

- The Fourier transform is periodic, with period \(N\). This is always true.

![Image of Gaussian and Fourier transform of Gaussian](http://www.cim.mcgill.ca/~langer/546/MATLAB/plotFourierTransformGaussian.jpg)

**Example 8: Gabor**

To compute the Fourier transform of Gabor, we use a property which is similar to the convolution theorem:

\[
\mathbf{F}(I(x)h(x))) = \frac{1}{N} \mathbf{F}I(x) \ast \mathbf{F}h(x).
\]

See Appendix B for a proof, if you are interested (not on exam).

Thus the Fourier transform of a cosine Gabor is the convolution in the frequency domain of the Fourier transforms of a Gaussian and the Fourier transform of a cosine:

\[
\mathbf{F} \cosGabor(x, k_0, \sigma) = \mathbf{F} \{ G(x, \sigma) \cos\left(\frac{2\pi}{N} k_0 x \right) \} \\
= \frac{1}{N} e^{-\frac{1}{2} \left(\frac{2\pi}{N} k_0 \right)^2} \ast \frac{N}{2} \left( \delta(k_0 - k) + \delta(k_0 + k) \right) \\
= \frac{1}{2} \left\{ e^{-\frac{1}{2} \left(\frac{2\pi}{N} (k-k_0) \right)^2} + e^{-\frac{1}{2} \left(\frac{2\pi}{N} (k+k_0) \right)^2} \right\}
\]

which is the sum of two Gaussians, centered at \(k = \pm k_0\).

**[ADDED: April 12, 2018]**

An example is shown below which was computed using Matlab. The cosine Gabor is defined on a vector of size \(N = 128\) and has a central frequency of 20 cycles and a Gaussian with a standard deviation of 5. The amplitude spectrum has a peak at \(k_0 = \pm 20\). In the amplitude spectrum plot on the right below, I plot the frequency range from 0 to \(N - 1\) instead of \(-\frac{N}{2}\) to \(\frac{N}{2} + 1\). The Fourier transform of a sine Gabor can be calculated similarly. (See Exercises.)
The convolution theorem tells us that convolving a function $I(x)$ with a cosine (or sine) Gabor will give you a function that has only a band of frequencies remaining, namely the frequencies near the center frequency $k_0$ of the Gabor. The width of the band depends on the $\sigma$ of the Gaussian of the Gabor. We will return to this idea of filtering a signal into bands of different frequencies (different Gabor filters can be used, or other bandpass filters) when we discuss audition.
Appendix A

We will use the following claim to show what the inverse Fourier transform is (bottom of page).

Claim (Example 4): For any frequency $k_0$,

$$F e^{i \frac{2\pi}{N} k_0 x} = N \delta(k - k_0).$$

That is,

$$\sum_{x=0}^{N-1} e^{i \frac{2\pi}{N} k_0 x} e^{-i \frac{2\pi}{N} k x} = \begin{cases} N, & k = k_0 \\ 0, & k \neq k_0 \end{cases}$$

Note that this claim essentially is essentially equivalent to saying that two cosine (or sine) functions of different frequencies are orthogonal; their inner product is 0.

Proof: Rewrite the left side of the above summation as

$$\sum_{x=0}^{N-1} e^{i \frac{2\pi}{N} (k_0-k)x}.$$  \hfill (2)

If $k = k_0$, then the exponent is 0 and so we are just summing $e^0 = 1$ and the result is $N$.

That doesn’t yet give us the result of the claim, because we still need to show that the summation is 0 when $k \neq k_0$. So, for the case $k \neq k_0$, observe that the summation is a finite geometric series and thus we can use the following identity which you know from Calculus\(^2\); let $\gamma$ be any number (real or complex) then

$$\sum_{x=0}^{N-1} \gamma^x = \frac{1 - \gamma^N}{1 - \gamma}.$$  \hfill (3)

Applying this identity for our case, namely $\gamma = e^{i \frac{2\pi}{N} (k-k_0)}$, lets us write (2) as

$$\sum_{x=0}^{N-1} e^{i \frac{2\pi}{N} (k-k_0)x} = \frac{1 - e^{i \frac{2\pi}{N} (k-k_0)}}{1 - e^{i \frac{2\pi}{N} (k-k_0)}}.$$  \hfill (3)

The numerator on the right hand side vanishes because $k - k_0$ is an integer and so

$$e^{i 2\pi (k-k_0)} = 1.$$  \hfill (3)

What about the denominator? Since $k$ and $k_0$ are both in $0, \ldots, N-1$ and since we are considering the case that $k \neq k_0$, we know that $|k - k_0| < N$ and so $e^{i \frac{2\pi}{N} (k-k_0)} \neq 1$. Hence the denominator does not vanish. Since the numerator is 0 but the denominator is not 0, we can conclude that the right side of Eq. (3) is 0. Thus, the summation of (2) is 0, and so $F e^{i \frac{2\pi}{N} k_0 x} = 0$ when $k \neq k_0$. This completes the derivation for the case $k \neq k_0$.

Claim (inverse Fourier transform): $F^{-1} = \frac{1}{N} \bar{F}$

Proof:

The matrix $\frac{1}{N} \bar{F} \bar{F}$ is $N \times N$. The above example says that row $k_0$ and column $k$ of this matrix is 1 when $k_0 = k$ and 0 when $k_0 \neq k$ and hence this matrix is the unit diagonal.

---

\(^2\)If you are unsure where this comes from, see equations (1)-(6) of [GeometricSeries.html](http://mathworld.wolfram.com/GeometricSeries.html)
Appendix B: Conjugacy property of the Fourier transform

Claim: If $I(x)$ is a real valued function, then

$$\overline{\hat{I}(k)} = \hat{I}(N - k).$$

Proof: (not on final exam)

$$\hat{I}(N - k) = \sum_{x=0}^{N-1} I(x) e^{-i \frac{2\pi}{N} (N-k)x}$$
$$= \sum_{x=0}^{N-1} I(x) e^{-i \frac{2\pi}{N} kx} e^{i \frac{2\pi}{N} x}$$
$$= \sum_{x=0}^{N-1} I(x) e^{i \frac{2\pi}{N} kx}, \text{ since } e^{i2\pi x} = 1 \text{ for any integer } x$$
$$= \sum_{x=0}^{N-1} I(x) \overline{e^{-i \frac{2\pi}{N} kx}}$$
$$= \sum_{x=0}^{N-1} \overline{I(x)} e^{i \frac{2\pi}{N} kx}, \text{ if } I(x) \text{ is real}$$
$$= \sum_{x=0}^{N-1} \overline{I(x)} e^{-i \frac{2\pi}{N} kx}$$
$$= \overline{\hat{I}(k)}$$
Appendix C: Convolution Theorem

Claim: For any two functions \( I(x) \) and \( h(x) \) that are defined on \( N \) consecutive samples e.g. 0 to \( N - 1 \),

\[
F(I(x) * h(x))) = F(I(x)) \cdot F(h(x)) = \hat{I}(k) \cdot \hat{h}(k).
\]

Proof: (not on final exam)

\[
\begin{align*}
F \cdot I * h(x) &= \sum_{x=0}^{N-1} e^{-\frac{i 2 \pi k x}{N}} \sum_{x'=0}^{N-1} I(x - x') h(x'), \text{ by definition} \\
&= \sum_{x'=0}^{N-1} h(x') \sum_{x=0}^{N-1} e^{-\frac{i 2 \pi k x}{N}} I(x - x'), \text{ by switching order of sums} \\
&= \sum_{x'=0}^{N-1} h(x') \sum_{u=0}^{N-1} e^{-\frac{i 2 \pi k (u+x')}{N}} I(u), \text{ where } u = x - x' \\
&= \sum_{x'=0}^{N-1} h(x') e^{-\frac{i 2 \pi k x'}{N}} \sum_{u=0}^{N-1} e^{-\frac{i 2 \pi k u}{N}} I(u) \\
&= \hat{h}(k) \hat{I}(k)
\end{align*}
\]
Appendix D (another Convolution Theorem)

We will often work with filters such as Gabor functions that are the product of two functions. Suppose we have two 1D functions \( I(x) \) and \( h(x) \) and we take their product. What can we say about the Fourier transform? The answer is similar to the convolution theorem, and indeed is just another version of that theorem:

\[
\mathcal{F} \left( I(x)h(x) \right) = \frac{1}{N} \hat{I}(k) \ast \hat{h}(k)
\]

or, in words, the Fourier transform of the product of two functions is the convolution of the Fourier transforms of the two functions. Note that the convolution on the right hand side is between two complex valued functions, rather than real valued functions. But the same definition of convolution applies.

To prove the above property, we take the inverse Fourier transform of the right side and show that it gives \( I(x)h(x) \). Note that the summations and functions below are defined on frequencies \( k, k', k'' \mod N \), since the Fourier transform of a function has period \( N \).

\[
\mathcal{F}^{-1} \hat{I}(k) \ast \hat{h}(k) = \frac{1}{N} \sum_{k=0}^{N-1} e^{\frac{2\pi}{N} kx} \sum_{k'=0}^{N-1} \hat{h}(k') \hat{I}(k - k') \quad \text{...and rearrange...}
\]

\[
= \frac{1}{N} \sum_{k'=0}^{N-1} \hat{h}(k') e^{\frac{2\pi}{N} k'x} \sum_{k=0}^{N-1} e^{\frac{2\pi}{N} (k-k')x} \hat{I}(k - k')
\]

\[
= \frac{1}{N} \sum_{k'=0}^{N-1} \hat{h}(k') e^{\frac{2\pi}{N} k'x} (\sum_{k=0}^{N-1} e^{\frac{2\pi}{N} (k-k'')x} \hat{I}(k - k''))
\]

\[
= \left\{ \begin{array}{l}
h(x) (\sum_{k''=-k'}^{N-1-k'} e^{\frac{2\pi}{N} (k'')x}) \hat{I}(k''), \quad \text{where } k'' = k - k' \\
Nh(x) I(x)
\end{array} \right.
\]

Dividing both sides by \( N \) and we’re done.

Sound waves

Last lecture we considered sound to be a pressure function \( I(X,Y,Z,t) \). However, sound is not just any function of those four variables. Rather, sound obeys the wave equation:

\[
\frac{\partial^2 I(X,Y,Z,t)}{\partial X^2} + \frac{\partial^2 I(X,Y,Z,t)}{\partial Y^2} + \frac{\partial^2 I(X,Y,Z,t)}{\partial Z^2} = \frac{1}{v^2} \frac{\partial^2 I(X,Y,Z,t)}{\partial t^2}
\]

where \( v \) is the speed of sound. This equation says that if you take a snapshot of the pressure function at any time \( t \), then the spatial derivatives the pressure function at each point \( XYZ \) tell you how the pressure at the point will change as time varies. Note that this equation contains the constant \( v \) which is the speed of sound.

The speed of sound in air is about \( v = 340 \) meters per second, or 34 cm per millisecond. This is quite slow. (If you go to a baseball game and you sit behind the outfield fence over 100 m away,
you can easily perceive the delay between when you see the ball hit the bat, and when you hear the ball hit the bat.) Amazingly, the speed of sound is so slow that our brains can detect differences in the arrival times of sounds at the left and right ear, and we use this difference to help us perceive where sound sources are. (We’ll discuss this in the following few lectures.)

Also notice that the wave equation is linear in \( I(X,Y,Z,t) \). If you have several sources of sound, then the pressure function \( I \) that results is identical to the sum of the pressure functions produced by the individual sources in isolation.

Today we will examine two types of sounds that are of great interest: music and speech. We will see how a frequency domain analysis is fundamental to both.

**Musical sounds**

Let’s begin by briefly considering string instruments such as guitars. First consider the vibrating string. When we pluck the guitar string, we are setting its initial shape to something different than its resting state. This initial shape and the subsequent shape as it vibrates always has fixed end points. The initial shape can be written as a sum of sine functions, specifically sine functions with value zero at the end points. This is summation is similar a Fourier transform, but here we only need sine functions (not sines and cosines), in particular,

\[
\sin\left(\frac{\pi}{L}x_m\right)
\]

where \( m \geq 0 \) is an integer and \( L \) is the length of the guitar string. We have \( \pi \) rather than \( 2\pi \) in the numerator since the sine value is zero when \( x = \frac{L}{m} \) for any \( m \).

Physics tells us that if a string is of length \( L \) then its mode \( \sin\left(\frac{\pi}{L}x\right) \) vibrates at a temporal frequency \( \omega = \frac{c}{L} \) where \( c \) is a constant that depends on the properties of the string such as its material, thickness, tension. Think of each mode \( m \) of vibration as dividing the string into equal size parts of size \( \frac{L}{m} \). For example, we would have four parts of length \( \frac{L}{4} \). (See sketch in slide). You can think of each of these parts as being little strings with fixed endpoints.

Frequency \( m \) is called the \( m \)-th harmonic. The frequency \( \omega_0 = \frac{c}{L} \) i.e. \( m = 1 \) is called the fundamental frequency. Frequencies for \( m > 1 \) are called overtones. Note harmonic frequencies have a linear progression \( m\omega_0 \). They are multiples of the fundamental.
Note that the *definition* of harmonic frequencies is that they are an integer multiple of a fundamental frequency. It just happens to be the case that vibrating strings naturally produce a set of harmonic frequencies. There are other ways to get harmonic frequencies as well, for example, voiced sounds as we will see later.

For stringed instruments such as a guitar, most of the sound that you hear comes not from the vibrating strings, but rather the sound comes from the vibrations of the instrument body (neck, front and back plates) in response to the vibrating strings. The body has its own vibration modes as shown below. The curved lines in the figure are the nodal points which do not move. Unlike the string, the body modes do not define an arithmetic progression.

For another example, see [http://www.acs.psu.edu/drussell/guitars/hummingbird.html](http://www.acs.psu.edu/drussell/guitars/hummingbird.html)

In western music, notes have letter names and are periodic: A, B, C, D, E, F, G, A, B, C, D, E, F, G, A, B, C, D, E, F, G, etc. Each of these notes defines a fundamental frequency. The consecutive fundamental frequencies of the notes for any letter (say C) are separated by one octave. e.g. A, B, C, D, E, F, G, A covers one octave. Recall from the linear systems lecture that a difference of one octave is a doubling of the frequency, and in general two frequencies $\omega_1$ and $\omega_2$ are separated by $\log_2 \frac{\omega_2}{\omega_1}$ octaves.

An octave is partitioned into 12 intervals called *semitones*. The intervals are each $\frac{1}{12}$ of an octave, i.e. equal intervals on a log scale. A to B, C to D, D to E, F to G, and G to A are all two semitones, whereas B to C and E to F are each one semitone. (No, I don’t know the history of that.) It follows that the number of semitones between a note with fundamental $\omega_1$ and a note with fundamental $\omega_2$ is $12 \log_2 \frac{\omega_2}{\omega_1}$. To put it another way, the frequency that is $n$ semitones above $\omega_1$ is $\omega_1 2^{\frac{n}{12}}$. The notes on a piano keyboard are shown below, along with a plot of their fundamental frequencies.
Notice that the frequencies of consecutive semitones define a geometric progression, whereas consecutive harmonics of a string define an arithmetic progression. When you play a note on a piano keyboard, the sound that results contains the fundamental as well as all the overtones - which form an arithmetic progression. When you play multiple notes, the sound contains the fundamentals of each note as well as the overtones of each. [ASIDE: The reason why some chords (multiple notes played together) sound better than other has to do – in part – with the distribution of the overtones of the notes, namely how well they align. Details omitted.]

Speech sounds

Human speech sounds have very particular properties. They obey certain physical constraints, namely our anatomy. Speech sounds depend on several variables. One is the shape of the oral cavity, which is the space inside your mouth. This shape is defined by the tongue, lips, and jaw position which are known as articulators. The sound wave that you hear has passed from the lungs, past the vocal cords, and through the long cavity (pharynx + oral and nasal cavity) before it exits the body. The shape of the oral cavity is determined by the position of the tongue, the jaw, the lips.

Consider the different vowel sounds in normal spoken English “aaaaaa”, “eeeeeee”, “iiiiiii”, “oooooo”, “uuuuuu”. Make these sounds to yourself and notice how you need to move your tongue, lips, and jaw around. These variations are determined by the positioning of the articulators. Think of the vocal tract (the volume between the vocal cords and the mouth and nose) as a resonant tube, like a bottle. Changing the shape of the tube by varying the articulators causes different sound frequencies that are emitted from you to be amplified and others to be attenuated.

Voiced Sounds

Certain sounds require that your vocal cords vibrate while other sounds require that they do not vibrate. When vocal cords are tensed, the sounds that result are called voiced. An example is a tone produced by a singing voice. When the vocal cords are relaxed, the sounds are called unvoiced. An example is whispering. Normal human speech is a combination of voiced and unvoiced sounds.

Voiced sounds are formed by regular pulses of air from the vocal cords. There is an opening in the vocal cords called the glottis. When the vocal cords are tensed, the glottis opens and closes at a regular rate. A typical rate for glottal “pulses” for adult males and females are around 100 and 200 Hz i.e. about a 10 ms or 5 ms period, although this can vary a lot depending on whether one has...
a deep versus average versus high voice. Moreover, each person can change their glottal frequency by varying the tension. That is what happens when you sing different notes.

Suppose you have \( n_{\text{glottal}} \) glottal pulses which occur with period \( T_{\text{glottal}} \) (time between pulses). The total duration would be \( T = n_{\text{glottal}} T_{\text{glottal}} \) time samples. We can write the sound source pressure signal that is due to the glottal pulse train as:

\[
I(t) = \sum_{j=0}^{n_{\text{glottal}}-1} g(t - jT_{\text{glottal}})
\]

where \( g() \) is the sound pressure due to each glottal pulse. We can write this equivalently as

\[
I(t) = g(t) * \sum_{j=0}^{n_{\text{glottal}}-1} \delta(t - jT_{\text{glottal}}).
\]

Each glottal pulse gets further shaped by the oral and nasal cavities. The oral cavity in particular depends on the positions of the articulators. If the articulators are fixed in place over some time interval, each glottal pulse will undergo the same waveform change in that interval. Some people speak very quickly but not so quickly that the position of the tongue, jaw and mouth changes over time scales of the order of say 10 ms. Indeed, if you could move your articulators that quickly, then your speech would not be comprehensible.

One can model the transformed glottal pulse train as a convolution with a function \( a(t) \), so the final emitted sound is:

\[
I(t) = a(t) * g(t) * \sum_{j=0}^{n_{\text{glottal}}-1} \delta(t - jT_{\text{glottal}})
\]

So you can think of \( a(t) * g(t) \) as a single impulse response function. The reason for separating them is that there really are two different things happening here. The glottal pulse \( g(t) \) is not an impulse function and it is different from the effect \( a(t) \) of the articulators. Each glottal pulse produces its own \( a(t) * g(t) \) pressure wave and these little waves follow one after the other.

Let’s next briefly consider the frequency properties of voiced sounds. If we take the Fourier transform of \( I(t) \) over \( T \) time samples – and we assume the articulators are fixed in position so that we can define \( a(t) \) and we assume \( T_{\text{glottal}} \) is fixed over that time also – we get

\[
\hat{I}(\omega) = \hat{a}(\omega) \hat{g}(\omega) F \sum_{j=0}^{n_{\text{glottal}}-1} \delta(t - jT_{\text{glottal}}).
\]

You can show (in Assignment 3) that

\[
F \sum_{j=0}^{n_{\text{glottal}}} \delta(t - jT_{\text{glottal}}) = n_{\text{glottal}} \sum_{j=0}^{T_{\text{glottal}}-1} \delta(\omega - jn_{\text{glottal}})
\]

So,

\[
\hat{I}(\omega) = \hat{a}(\omega) \hat{g}(\omega) n_{\text{glottal}} \sum_{j=0}^{T_{\text{glottal}}-1} \delta(\omega - jn_{\text{glottal}})
\]
This means that the glottal pulses cancel out all frequencies except other than those that are a multiple of \( n_{glottal} = \frac{T}{T_{glottal}} \), that is, the number glottal pulses per \( T \) samples. I emphasize here that this clean mathematical result requires that the sequence of glottal pulses spans the \( T \) samples, and the period is regular and the articulators are fixed during that interval.

Measurements show that the glottal pulse \( g(t) \) is a low pass function. You can think of it as having a smooth amplitude spectrum, somewhere between a Gaussian amplitude spectrum which falls off quickly and an impulse amplitude spectrum which is constant over \( \omega \).

The effect of the articulators is to modulate the amplitude spectrum that is produced by the glottal pulses, namely by multiplying by \( \hat{a}(\omega) \). This amplifies some frequencies and attenuates others. (It also produces phase shifts which we will ignore in this analysis, but which are important if one considers the wave shape of each pulse.) The peaks of the amplitude spectrum \( |\hat{g}(\omega) \hat{a}(\omega)| \) are called formants. As you change the shape of your mouth and you move your jaw, you change \( a(t) \) which changes the frequencies of the formants. I will mention formants again later when I discuss spectrograms.

As mentioned above, the sum of delta functions nulls out frequencies except those that happen to be part of an arithmetic progression of fundamental frequency \( \omega_0 = n_{glottal} = \frac{T}{T_{glottal}} \), that is, \( n_{glottal} \) samples per \( T \) time steps. However, we often want to express our frequencies in cycles per second rather than cycles per \( T \) samples. The typical sampling rate used in high quality digital audio is 44,100 samples per second, or about 44 samples per ms.\(^3\) To convert from cycles per \( T \) samples to cycles per second, one should multiply by 44,100 samples per second.

This sampling rate is not the only one that is used, though. Telephone uses a lower sampling rate, for example, since quality is less important.

The frequency 44,100 * \( n_{glottal} \) is the fundamental frequency in cycles per second, which corresponds to the glottal pulse train. As mentioned earlier, in adult males this is typically around 100 Hz for normal spoken voice. In adult females, it is typically around 200 Hz. In children, it is often higher than 250 Hz.

The two rows in the figure below illustrate a voiced sound with fundamental 100 and 200 Hz. The left panels shows just amplitude spectrum of the glottal pulse train. The center panels illustrate the amplitude spectrum of the articulators for several formants. The right panel shows the amplitude spectrum of the resulting sound.

\(^3\)One often uses 16 bits for each of two channels (two speakers or two headphones).
Unvoiced sounds (whispering)

When the vocal cords are relaxed, the resulting sounds are called unvoiced. There are no glottal pulses. Instead, the sound wave that enters the oral cavity can be described better as noise. The changes that are produced by the articulators, etc are roughly the same in voiced versus unvoiced speech, but the sounds that are produced are quite different. You can still recognize speech when someone whispers. That’s because there is still the same shaping of the different frequencies into the formants, and so the vowels are still defined. But now it is the noise that gets shaped rather than glottal pulses.

I mentioned in the lecture that the noise \( n(t) \) produced by expelling air from the lungs has a flat amplitude spectrum, hat is, prior to the reshaping of the spectrum by the articulators. The sound that comes out the mouth is \( n(t) \ast a(t) \) and that sound is shaped by the articulators.

Consonants

Another important speech sound occurs when one restricts the flow of air, and force it through a small opening. For example, consider the sound produced when the upper front teeth contact the lower lip. Compare this to when the lower front teeth are put in contact with the upper lip. (The latter is not part of English. I suggest you amuse yourself by experimenting with the sounds you can make in this way.) Compare these to when the tongue is put in contact with the front part of the palate vs. the back part of the palate.

Most consonants are defined this way, namely by a partial or complete blockage of air flow. There are several classes of consonants. Let’s consider a few of them. For each, you should consider what is causing the blockage (lips, tongue, palate).

- fricatives (narrow constriction in vocal tract):
  - voiced: z, v, zh, th (as in the)
  - unvoiced: s, f, sh, th (as in θ)
• stops (temporary cessation of air flow):
  – voiced: b, d, g
  – unvoiced: p, t, k

These are distinguished by where in the mouth the flow is cutoff. Stops are accompanied by a brief silence

• nasals (oral cavity is blocked, but nasal cavity is open)
  – voiced: m, n, ng

You might not believe me when I tell you that nasal sounds actually come out of your nose. Try shutting your mouth, plugging your nose with your fingers, and saying ”mmmmmm”. See what happens?

Spectrograms

When we considered voiced sounds, we took the Fourier transform over $T$ samples and assumed that the voiced sound extended over those samples. One typically does not know in advance the duration of voiced sounds, so one has to arbitrary choose a time interval.

Often one analyzes the frequency content of a sound by partitioning $I(t)$ into blocks of $B$ disjoint intervals each containing $T$ samples – the total duration of the sound would be $BT$. For example, if $T = 512$ and the sampling rate is 44000 samples per second, then each interval would be about 12 milliseconds.

Let’s compute the discrete Fourier transform on the $T$ samples in each of these block. Let $\omega$ be the frequency variable, namely cycles per $T$ samples, where $\omega = 0, 1, \ldots, T - 1$. Consider a 2D function which is the Fourier transform of block $b$:

$$\hat{I}(b, \omega) = \sum_{t=0}^{T-1} I( b T + t) e^{-i \frac{2\pi}{T} \omega t}.$$ 

Typically one ignores the phase of the Fourier transform here, and so one only plots the amplitude $|\hat{I}(b, \omega)|$. You can plot such a function as a 2D “image”, which is called a spectrogram.

The sketch in the middle shows a spectrogram with a smaller $T$, and the sketch on the right shows one with a larger $T$. The one in the middle is called a ”wideband” spectrogram because each ’pixel’ of the spectrogram has a wide range of frequencies, and the one on the right is called a narrowband spectrogram because each ’pixel’ has a smaller range of frequencies. For example, if $T = 512$ samples, each pixel would be about 12 ms wide and the steps in $\omega$ would be 86 Hz high, whereas if $T = 2048$ samples, then each pixel would be be 48 ms wide and the $\omega$ steps would be 21 Hz.

Notice that we cannot simultaneously localize the properties of the signal in time and in frequency. If you want good frequency resolution (small $\omega$ steps), then you need to estimate the frequency components over long time intervals. Similarly, if you want good temporal resolution (i.e. when exactly does something happen?), then you can only make coarse statements about which frequencies are present ”when” that event happens. This inverse relationship is similar to what we observed earlier when we discussed the Gaussian and its Fourier transform.
Examples (see slides)

The slides show a few examples of spectrograms of speech sounds, in particular, vowels. The horizontal bands of frequencies are the formants which I mentioned earlier. Each vowel sound is characterized by the relative positions of the three formants. For an adult male, the first formant (called F1) is typically centered anywhere from 200 to 800 Hz. The second formant F2 from 800 to 2400 Hz, F3 from 2000 to 3000 Hz.