[MUSIC PLAYING]

PROFESSOR: Last time, we began the discussion of the z-transform. As with the Laplace transform in continuous time, we developed it as a generalization of the Fourier transform. The expression that we got for the z-transform is the sum that I indicate here. We also briefly talked about an inverse z-transform integral and some other informal methods of computing the inverse z-transform. But we focused in particular on the relationship between the z-transform and the Fourier transform, pointing out first of all that the z-transform, when we choose the magnitude of z equal to 1-- so the magnitude of z of the form e to the j omega-- just simply reduces to the Fourier transform of the sequence.

Then, in addition, we explored the z-transform for z, a more general complex number. In the discrete time z-transform case, we expressed that complex number in polar form as r e to the j omega, and recognize that the z-transform expression in fact corresponds to the Fourier transform of the sequence exponentially weighted. Because of the exponential weighting, the z-transform converges for some values of r corresponding to some exponential waiting, and perhaps not for others, and that led to a notion which corresponded to the region of convergence associated with the z-transform. We talked some about properties of the region of convergence, particularly in relation to the pole-zero pattern.

Now, they z-transform has a number of important and useful properties, just as the Laplace transform does. As one part of this lecture, what we'll want to do is exploit some of these properties in the context of systems described by linear constant coefficient difference equations. These particular properties that play an important role in that context are the properties that I indicate here. In particular, there is-- as

with continuous time-- a linearity property that tells us that the z-transform of a linear combination of sequences is the same linear combination of the z-transforms, a shifting property that indicates that the z-transform of x of n shifted is the z transform of x of n multiplied by a factor z to the minus n 0. Then the convolution property for which the z-transform of a convolution of sequences is the product of the associated z-transforms.

With all of these properties, of course, there is again the issue of what the associated region of convergence is in comparison with the region of convergence of the original sequences. That is an issue that's addressed somewhat more in the text and let's not go into it here.

With the convolution property, the convolution property as in continuous time, of course, provides a mechanism for dealing with linear time invariant systems. In particular, in the time domain a linear time invariant system is described through convolution-- namely, the output is the convolution of the input and the impulse response. Because of the convolution property associated with the z-transform, the z-transform of the output is the z-transform of the input times the z-transform of the impulse response. Again, very much the same as what we had in continuous time and also what we had in the context of the discussion with the Fourier transform.

In fact, because of the relationship between the z-transform and the Fourier transform, the z-transform of the impulse response evaluated on the unit circle-- in other words, for the magnitude of z equal to 1-- in fact corresponds to the frequency response of the system. More generally, when we talk about the z-transform of the impulse response, we will refer to it as the system function associated with the system.

Now, the convolution property and these other properties, as I indicated, we will find useful in talking about systems which are described by linear constant coefficient difference equations, and in fact, we'll do that shortly. But first what I'd like to do is to continue to focus on the system function for linear time invariant systems, and make a couple of comments that tie back to some things that we said in the last lecture

relating to the relationship between the region of convergence of a system, or-- I'm sorry, the region of convergence of a z-transform-- and the issue of where that is in relation to the poles of the z-transform.

In particular, we can draw some conclusions tying back to that discussion about the pole locations of the system function in relation to whether the system is stable and whether the system is causal. In particular, recall from one of the early lectures way back when that stability for a system corresponded to the statement that the impulse response is absolutely summable. Furthermore, when we talked about the Fourier transform, the Fourier transform of a sequence converges if the sequence is absolutely summable. So in fact, the condition for stability of a system and the condition for convergence of the Fourier transform of its impulse response are the same condition, namely absolute summability.

Now what does this mean? What it means is that if the Fourier transform converges, that means that the z-transform converges on the unit circle. Consequently, if the system is stable, then the system function, the z-transform or the impulse response, must also converge on the unit circle. In other words, the impulse response must have a Fourier transform that converges. So for a stable system, then, the region of convergence of the system function must include the unit circle in the z-plane.

So we see how stability relates to the location of the region of convergence, and we can also relate causality to the region of convergence. In particular, we know that if a system is causal, then the impulse response is right-sided. For a sequence that's right-sided, the region of convergence of its z-transform must be outside the outermost pole. So for causality, the region of convergence of the system function must be outside the outermost pole. For stability, the region of convergence must include the unit circle, and we can also then draw from that the conclusion that if we have a system that's causal unstable, then all poles must be inside the unit circle because of the fact that the poles must-- because of the fact that the region of convergence must be outside the outermost pole and has to also include the unit circle.

So for example, if we had let's say a system with a system function as I indicate here with the algebraic expression for the system function being the expression that I indicate here with the pole at z equals a third and another pole at z equals 2 and a zero at the origin. If, in fact, the system was causal, corresponding to an impulse response that's right-sided, this then would be the region of convergence of the system function.

Alternatively, if I knew that the system was stable, then I know that the region of convergence must include the unit circle, and so this would be then the region of convergence. And what you might now want to ask yourselves is if instead the region of convergence for the system function is this, then is the system causal? That's the first question. The second question is is the system stable? Remembering that for causality, the region of convergence must be outside the outermost pole, and for stability it must include the unit circle.

Now what I'd like to do is look at the properties of the z-transform, and in particular exploit these properties in the context of systems that are described by linear constant coefficient difference equations. The three basic properties that play a key role in that discussion are the linearity property, the shifting property, and the convolution property. These, then, are the properties that I want to exploit.

So let's do that by first looking at a first order difference equation. In the case of a first order difference equation, which I've written as I indicate here-- no terms on the right hand side, but we could have terms of course, in general-- y of n minus ay of n minus 1 equals x of n. We can use the linearity property, so that if we take the z-transform of both sides of this expression, that will then be the z-transform of this term plus the z-transform of this term. So using those properties and together with the shifting property, the property that tells us that the z-transform of y of n minus 1 is z to the minus 1 times the z-transform of y of n, we then convert the difference equation to an algebraic expression. And we can solve this algebraic expression for the z-transform of the output in terms of the z-transform of the input.

Now what we know from the convolution property is that for a system, the z-

transform of the output is the system function times the z-transform of the input. So this factor that we have, then, must correspond to the system function, or equivalently the z-transform of the impulse response of the system. In fact, then, if we have this z-transform, we could figure out what the impulse response of the system is by computing or determining what the inverse z-transform, except for the fact that expression is an algebraic expression and doesn't yet totally specify the z-transform because we don't yet know what the region of convergence is.

How do we get the region of convergence? Well, we have the same issue here as we had with the Laplace transform-- namely, the point that the difference equation tells us, in essence, what the algebraic expression is for the system function, but doesn't specify the region of convergence. That is specified by either explicitly, because one way or another we know what the impulse response is, or implicitly, because we know certain properties of the system, such as causality and or stability.

If I, let's say, imposed on this system, in addition to the difference equation, the condition of causality, then what that requires is that the impulse response be right-sided or the region of convergence be outside the outermost pole. So, for this example that would require, then, that the region of convergence correspond to the magnitude of z greater than the magnitude of a.

What you might think about is whether I would get the same condition if I required instead that the system be stable. Furthermore, you could think about the question of whether I could specify or impose on this system that it be both stable and causal. The real issue-- and let me just kind of point to it-- is that the answer to those questions relate to whether the magnitude of a is less than 1 or the magnitude of a is greater than 1. If the magnitude of a is less than 1 and I specify causality, that will also mean that the system is stable. In any case, given this region of convergence, then the impulse response is the inverse transform of that z-transform, which is a to the n times u of n.

Now let's look at a second order equation, and there's a very similar strategy. For

the second order equation, the one that I've picked is of this particular form, and I've written the coefficients parametrically as I indicate here for a specific reason, which we'll see shortly. Again, I can apply the z-transform to this expression, and I've skipped an algebraic step or two here. When I do this, then, again I use the linearity property and the shifting property, and I end up with this algebraic expression. If I now solve that to express y of z in terms of x of z and a function of z, then this is what I get for the system function. So this is now a second order system function, and we'll have two zeroes at the origin and it will have two poles.

Again, there's the question of what we assume about the region of convergence-- I haven't specified that yet. But if we, let's say, assume that the system is causal, which I will tend to do, then that means that the region of convergence is outside the outermost pole.

Now, where are the poles? Well, let me just kind of indicate that if-- and you can verify this algebraically at your leisure-- that if the cosine theta term is less than 1, then the roots of this polynomial will be complex. And in fact, the poles are at r e to the plus or minus j theta. So for cosine theta less than 1, then the poles are complex, and the complex poles at an angle theta and with a distance from the origin equal to the parameter r.

In fact, let's just look at that. What I show here is the pole zero pattern. Assuming that r is less than 1, and that cosine theta is less than 1, and we have a complex pole pair shown here-- now if we assume that the system was causal, that means that the region of convergence is outside these poles. That would then include the unit circle, which means that the system is also stable. In fact, as long as the reach of convergence includes the unit circle, we can also talk about the frequency response of the system-- namely, we can evaluate the system function on the unit circle.

We, in fact, evaluated the Fourier transform associated with this pole zero pattern last time. Recall that the frequency response, then, is one that has a resonant character with the resonant peak being roughly in the vicinity of the angle of the

pole location. As the parameter r varies-- let's say, as r gets smaller-- this peak tends to broaden. As r gets closer to 1, the resonance tends to get sharper.

This is now a look at the z-transform, and we see very strong parallels to the Laplace transform. In fact, throughout the course, I've tried to emphasize-- and it just naturally happens-- that there are very strong relationships and parallels between continuous time and discrete time. In fact, at one point we specifically mapped from continuous time to discrete time when we talked about discrete time processing of continuous time signals. What I'd like to do now is turn our attention to another very important reason for mapping from continuous time to discrete time, and in the process of doing this, what we'll need to do is exploit fairly heavily the insight, intuition, and procedures that we've developed for the Laplace transform and the z-transform.

Specifically, what I would like to begin is a discussion relating to mapping continuous time filters to discrete time filters, or continuous time system functions to discrete time system functions. Now, why would we want to do that? Well, there are at least several reasons for wanting to map continuous time filters to discrete time filters. One, of course, is the fact that in some situations what we're interested in doing is processing continuous time signals with discrete time systems-- or, said another way, simulate continuous time systems with discrete time systems. So it would be natural in a setting like that to think of mapping the desired continuous time filter to a discrete time filter.

So that's one very important context. There's another very important context in which this is done, and that is in the context or for the purpose of exploiting established design procedures for continuous time filters. The point is the following. We may or may not be processing a sample continuous time signal with our discrete time filter-- it may just be discrete time signals that we're working with. But in that situation, still, we need to design the appropriate discrete time filter.

Historically, there is a very rich history associated with design of continuous time filters. In many cases, it's possible and very worthwhile and efficient to take those

designs and map them to discrete time designs to use them as discrete time filters. So, another very important reason for talking about the kinds of mappings that we will be going into is to simply take advantage of what has been done historically in the continuous time case.

Now, if we want to map continuous time filters to discrete time filters, then in continuous time, we're talking about a system function and an associated differential equation. In discrete time, there is the corresponding system function and the corresponding difference equation. Basically what we want to do is generate from a continuous time system in some way a discrete time system that meets an associated set of desired specifications.

Now, there are certain constraints that it's reasonable and important to impose on whatever kinds of mappings we use. Obviously, we want a mapping that will take our continuous time system function and map it to a discrete time system function. Correspondingly in the time domain, there is a continuous time impulse response that maps to the associated discrete time impulse response. These are more or less natural. The two that are important and sometimes easy to lose sight of are the two that I indicate here.

In particular, if we are mapping a continuous time filter with, let's say, a desired or desirable frequency response to a discrete time filter and we would like to preserve the good qualities of that frequency response as we look at the discrete time frequency response, then it's important what happens in the s-plane for the continuous time filter along the j omega axis relate in a nice way to what happens in the z-plane around the unit circle, because it's this over here that represents the frequency response in continuous time and this contour over here that represents the frequency response in discrete time. So that's an important property. We want to kind of the j omega axis to map to the unit circle. Another more or less natural condition to impose is a condition that if we are assured in some way that our continuous time filter is stable, then we would like to concentrate on design procedures that more or less preserve that and will give us stable digital filters.

So these are kind of reasonable conditions to impose on the procedure. What I'd like to do in the remainder of this lecture is look at two common procedures for mapping continuous time filters to discrete time filters.

The first one that I want to talk about is one that, in fact, is very frequently used, and also one that, as we'll see for a variety of reasons, is highly undesirable. The second is one that is also frequently used, and as we'll see is, in certain situations, very desirable.

The first one that I want to talk about is the more or less intuitive simple procedure of mapping a differential equation to a difference equation by simply replacing derivatives by differences. The idea is that a derivative is more or less a difference, and there's some dummy parameter capital T that I've thrown in here, which I won't focus too much on.

But in any case, this seems to have some plausibility. If we take the differential equation and do this with all the derivatives, both in terms of y of t and x of t, what we'll end up with is a difference equation.

Now what we can use are the properties of the Laplace transform and the z-transform to see what this means in terms of a mapping-- in particular, using the differentiation property for Laplace transforms. In the Laplace transform domain, we would have this. Using the properties for the z-transform, the z-transform of this expression would be this. So, in effect, what it says is that every place in the system function or in the differential equation that we would be multiplying by s when Laplace transformed. In the difference equation, we would be multiplying by this factor.

In fact, what this means is that the mapping from continuous time to discrete time corresponds to taking the system function and replacing s wherever we see it by 1 minus z to the minus 1 over capital T. So if we have a system function in continuous time and we map it to a discrete time system function this way by replacing derivatives by differences, then that corresponds to replacing s by 1 minus z to the minus 1 over capital T.

Now we'll see shortly what this mapping actually means more specifically in relating the s-plane to the z-plane. Let me just quickly, because I want to refer to this, also point to another procedure very much like backward differences which corresponds to replacing derivatives not by the backward differences that I just showed, but by forward differences. In that case, then, the mapping corresponds to replacing s by z to the minus 1 over capital T. It looks very similar to the previous case. So there the relationship between these system functions is what I indicate here.

Let's just take a look at what those mappings correspond to when we look at this specifically in the s-plane and in the z-plane. What I show here is the s-plane, and of course it's things on the left half of the s-plane, poles on the left half of the s-plane that would guarantee stability. It's the j omega axis the tells us about the frequency response, and in the z-plane it's the unit circle that tells us about the frequency response. Things inside the unit circle, or poles inside the unit circle, that guarantee stability.

Now the mapping from s-plane to the z-plane corresponding to replacing derivatives by backward differences in fact can be shown to correspond to mapping the j omega axis not to the unit circle, but to the little circle that I show here, which is inside the unit circle. The left half of the s-plane maps to the inside of that circle. What does that mean?

That means that if we have a really good frequency response characteristic along this contour in the s-plane, we'll see that same frequency response along this little circle. That's not the one that we want, though-- we would like to see that same frequency response around the unit circle.

To emphasize this point even more-- suppose, for example, that we had a pair of poles in our continuous time system function that looked like this. Then, where they're likely to end up in the z-plane is inside the unit circle, of course. But if the poles here are close to the j omega axis, that means that these poles will be close to this circle, but in fact might be very far away from the unit circle. What would happen, then, is that if we saw in the continuous time filter a very sharp resonance,

the discrete time filter in fact might very well have that resonance broadened considerably because the poles are so far away from the unit circle.

Now, one plus with this method, and it's about the only one, is the fact that the left half of the s-plane maps inside the unit circle-- in fact, inside a circle inside the unit circle, and so stability is always guaranteed. Let me just quickly mention, and you'll have a chance to look at this a little more carefully in the video course manual, that for forward differences instead of backward differences, this contour in the s-plane maps to a line in the z-plane, which is a line tangent to the unit circle, and in fact is the line that I showed there. So not only are forward differences equally bad in terms of the issue of whether they map from the j omega axis to the unit circle, but they have a further difficulty associated with them, which is the difficulty they may not and generally don't guarantee stability.

Now, that's one method, and one, as I indicated, that's often used partly because it seems so intuitively plausible. What you can see is that by understanding carefully the issues and the techniques of Laplace and z-transforms, you can begin to see what some of the difficulties with those methods are.

The next method that I'd like to talk about is a method that, in fact, is very commonly used. It's a very important, useful method, which kind of can be motivated by thinking along the lines of mapping the continuous time filter to a discrete time filter in such a way that the shape of the impulse response is preserved-- and, in fact, more specifically so that the discrete time impulse response corresponds to samples of the continuous time impulse response. And this is a method that's referred to as impulse invariance.

So what impulse invariance corresponds to is designing the filter in such a way that the discrete time filter impulse response is simply a sample version of the continuous time filter impulse response with a sampling period which I denote here as capital T. That will turn into a slightly confusing parameter shortly, and perhaps carried over into the next lecture. Hopefully, we'll get that straighted out, though, within those two lectures.

Remembering the issues of sampling, the discrete time frequency response, then since the frequency responses the Fourier transform of the impulse response is related to the continuous time impulse response as I indicate here, what this says is that it is the superposition of replications of the continuous time frequency response, linearly scaled in frequency and shifted and added to each other. It's the same old sort of shifting, adding, or aliasing issue-- the same sampling issues that we've addressed before.

This equation will help us understand what the frequency response looks like. But in terms of an analytical procedure for mapping the continuous time system function to a discrete time system function, we can see that and develop it in the following way.

Let's consider the continuous time system function expanded in a partial fraction expansion. And just for convenience, I'm picking first order poles-- this can be generalized multiple order poles, and we won't do that here. The same basic strategy applies. If we expand this in a partial fraction expansion, and we look at the impulse response associated with this-- we know how to take the inverse of Laplace transform of this, where I'm just naturally assuming causality throughout the discussion-- the continuous time impulse response, then, is the sum of exponentials with these amplitudes and at these complex exponential frequencies.

Now, impulse invariance corresponds to sampling this, and so the discrete time impulse response is simply a sampled version of this. The A sub k, of course, carries down. We have the exponential-- we're sampling at t equals n capital T, and so we've replaced that here, and then the unit step to truncate things for negative time.

Let's manipulate this further, and eventually what we want to get is a relationship-- a mapping-- from the continuous time to the discrete time filter. We have this step, and we can rewrite that now as I show here, just simply taking this n outside, and we have e to the s sub k capital T. Now this is of the form the sum of terms like A sub k times is beta to the n. We can compute the z-transform of this, and the z-transform that we get I show here-- it's simply A sub k over 1 minus e to the s sub k

capital T z to the minus 1, simply carrying this term or this parameter down.

So we started with a continuous time system function, which was a sum of terms like A sub k over s minus s sub k-- the poles were at s sub k. We now have the discrete time filter in this form. Consequently, then, this procedure of impulse invariance corresponds to mapping the continuous time filter to a discrete time filter by mapping the poles in the continuous time filter. According to this mapping, the continuous time filter pole at s sub k gets mapped to a pole e to the s sub k capital T, and the coefficients A sub k are preserved. That, then, algebraically, is what the procedure of impulse invariance corresponds to.

Let's look at how we interpret some of this in the frequency domain. In particular, we have the expression that tells us how the discrete time frequency response is related to the continuous time frequency response. This is the expression that we had previously when we had talked about issues of sampling. So that means that we would form the discrete time frequency response by taking the continuous time 1, scaling it in frequency according to this parameter capital T, and then adding replications of that together.

So if this is the continuous time frequency response, just simply an ideal low-pass filter with a cutoff frequency of omega sub c, then the frequency scaling operation would keep the same basic shape but linearly scale the frequency axis so that we now have omega sub c times T. Then the discrete time frequency response would be a superposition of these added together at multiples of 2 pi in discrete time frequency. So that's what we have here-- so this is the discrete time frequency response.

This looks very nice-- it looks like impulse invariance will take the continuous time frequency response, just simply linearly scale the frequency axis, and incidentally periodically repeat it. We know that for an ideal low-pass filter, that looks just fine. In fact, for a band-limited frequency response, that looks just fine. But we know also that any time that we're sampling-- and here we're sampling the impulse response-- we have an effect in the frequency domain or the potential for an affect an effect

that we refer to as aliasing.

So in fact, if instead of the ideal low-pass filter we had taken a filter that was an approximation to a low-pass filter, then the corresponding frequency scale version would look as I've shown here. And now as we add these together, then what we will have is some potential for distortion corresponding to the fact that these replications overlap, and what that will lead to is aliasing.

So some things that we can say about impulse invariance is that we have an algebraic procedure-- and I'll illustrate with another example shortly-- for taking a continuous time system function and mapping it to a discrete time system function. It has a very nice property in terms of mapping, the mapping from the frequency axis in continuous time due to the unit circle-- namely, to a first approximation. As long as there's no aliasing, the mapping is just simply a linear scaling of the frequency axis, although there may be some aliasing. That means, of course, that this method can only be used if the frequency response that's being mapped, or if the system that's being mapped, has a frequency response that's approximately low-pass-- it has to be approximately band-limited. Then what we have is some of potential distortion, which comes about because of aliasing.

Also because of the mapping, the fact that poles at s sub k get mapped to poles at e to the s sub k capital T, if the analog or continuous time filter is stable-- meaning that the real part of s sub k is negative-- then the discrete time filter is guaranteed to be stable. In other words, the magnitude of z sub k will be guaranteed to be less than 1. I'm assuming, of course, in that discussion that we are always imposing causality on the systems.

To just look at the algebraic mapping a little more carefully, let's take a simple example. Here is an example of a system, a continuous time system, where I simply have a resident pole pair with an imaginary part along the imaginary axis of omega sub r and a real part of minus alpha. And so the associated system function then is just the expression which incorporates the two poles, and I've put in a scale factor of 2 omega sub r.

And now to design the discrete time filter using impulse invariance, you would carry out a partial fraction expansion of this, and that partial fraction expansion is shown below. We have a pole at minus alpha minus j omega r and at minus alpha plus j omega r. And to determine the discrete time filter based on impulse invariance, we would map the poles and preserve the coefficients a sub k, referred to as the residues. And so the discrete time filter that we would end up with as a system function, which I indicate here-- and we have, as I said, preserved the residue, and the pole is now at a to the minus alpha T, e to the minus j omega sub r T. That's one term, and the other term in the sum has a pole at the complex conjugate location.

If we were to add these two factors together, then what we would get is both poles and a 0 at the origin. In fact, then, the pole is defined by its angle, and this angle is e to the j omega sub r capital T, and by its radius, and this radius is e to the minus alpha capital T. Now we can look at the frequency response associated with that, and let's just do that. For the original continuous time frequency response, what we have is simply a resonant character, as I've shown here. And if we map this using impulse invariance, which we just did, the resulting frequency response that we get is the frequency response which I indicate. We see that that's basically identical to the continuous time frequency response, except for a linear scaling in the frequency axis, if you just compare the dimensions along which the frequency axis is shown except for one minor issue, which is particularly highlighted when we look at the frequency response at higher frequencies.

What's the reason why those two curves don't quite follow each other at higher frequencies? Well, the reason is aliasing. In other words, what's happened is that in the process so applying impulse invariance, the frequency response of the original continuous time filter is approximately preserved, except for some distortion, that distortion corresponding to aliasing.

Well, just for comparison, let's look at what would happen if we took the same example-- and we're not going to work it through here carefully. We're not work it through it all, not even not carefully. If we took the same example and mapped it to

a discrete time filter by replacing derivatives by backward differences, what happens in that case is that we get a frequency response that I indicate here. Notice that the resonance in the original continuous time filter is totally lost. In fact, basically the character of the continuous time frequency response is lost.

What's the reason? Well, the reason goes back to the picture that I drew before. The continuous time filter had a pair of resident poles close to the j omega axis. When those get mapped using backward differences, they end up close to this little circle that's inside the unit circle, but in fact for this example, are far away from the unit circle.

So far we have one useful technique for mapping continuous time filters to discrete time filters. In part to highlight some of the issues, I focused attention also on some not so useful methods-- namely, mapping derivatives to forward or backward differences. Next time what I would like to do is look at impulse invariance for another example-- namely, the design of a Butterworth filter, and I'll talk more specifically about what Butterworth filters are at the beginning of that lecture.

Then, in addition, what we'll introduce is another very useful technique, which has some difficulties which impulse invariance doesn't have, but avoids the principal difficulty that impulse invariance does have-- namely, aliasing. That method is referred to the bilinear transformation, which we will define and utilize next time. Thank you.