Programmable Embedded Systems (EE60098)
Homework 8

Submitted By
Pratyush Jaiswal
18EE35014

Fom Discrete-time Signal Processing, Book by Alan V. Oppenheim and Ronald W. Schafer, read from chapter 6, sections 6.7 to 6.10. And submit a two-page short note.

There are numerous ways to implement LTI Discrete-time systems. Still, instead of direct from structures using infinite precision arithmetic because of their varied behaviour when the coefficients are quantized using a fixed point structure. For quantization, we use various representations like fixed-point and floating-point, each having its own set of benefits and disadvantages owing to quantization, rounding and truncation and related mathematical properties are observed. Quantization may introduce non-linearity in the system and exhibit zero-input limit cycles introduced by quantization and overflow. Higher-order systems have much more variety regarding choices in structure choice to counter quantization effects than lower-order ones. Although we are always interested in implementations requiring minor hardware or software complexity, it is not possible always. Quantization noise is internal to the system and adds up at all stages of the performance, and some structure coefficients are more sensitive to perturbation, and some are less. Kaiser in 1966 showed for IIR systems, a slight change in coefficients can cause a significant shift in poles and zeroes for direct form. If the poles are zeros clustered together as in narrow BPF or narrow BW, LPF poles are sensitive to coefficient

quantization errors. An increase in poles/zeroes increases sensitivity too. Cascade and parallel forms are less susceptible to coefficient quantization than equivalent direct forms and are preferred for higher-order implementations. Moreover, in the natural state, a 16-bit performance being unstable shows other issues too. But in FIR systems, we are concerned with zeros because for causal FIR system, all poles are at z=0, and direct form structure is used here. Also, uniformity in the spread of zeroes helps in direct-form FIR implementation. Many stimulating effects of quantization are studied using linear additive noise approximations. Quantization noises are assumed to be a) broad sense stationary white sense processes, b) uniform amplitude distribution over one quantization interval, c) Uncorrelated to the quantizer input, and the assumptions have led to accurate predictions. As poles approach unit circle, noise variance increases, requiring longer words to maintain conflict at the prescribed level. Also, note that different structures have different noise variances. In IIR systems, there are overflow issues and quantization help here, but the SNR also goes down, so there is a tradeoff. It has been seen that as the system pole approaches the unit circle, SNR falls as noise increases due to amplification and the high gain of the input to be scaled down by quantization and hence quantization and overflow works in opposite to reduce performance. Overflow problem is also seen in FIR fixed-point realization, and some considerations come into play. Floating-point works very well with accuracy for small systems, but the fixed point is used for high volume systems. In IIR digital filters using finite register length arithmetic, the output may continue to oscillate in periodic pattern even when input is zero, which is zero

input limit cycle behaviour. When a recursive system enters a limit cycle, it can be brought out only with a higher input than its dead band. Overflow, in addition, causes it and can be eliminated using saturation arithmetic. FIR systems do not have this issue as there are no recursive implementations.