The scalability extension of H.264/AVC uses a layered MCTF approach for providing spatial, temporal, and SNR scalability. The underlying MCTF decomposition scheme generally results in a large structural encoding–decoding delay, which prevents the use in low–delay applications. In this paper we present a new strategy for restricting the MCTF decomposition in way that arbitrary maximum structural delays can be adjusted. Based on this approach, the impact of the encoding–decoding delay on the coding efficiency of the scalable extension of H.264/AVC is analyzed.