	D del (for 's holeful to so	ali anno anno da Phona				To this and this		
(Jross-modal trans	iter is helptul to er	nhance modality-s	pecific discrimina	tive power for sc	ene recognition.	To this end, this	paper presents a	unified framework to i