Recent studies have revealed a strong association between human speech production and perception: on one hand, the neural system relies on the perception of the acoustic signal to generate clear speech; on the other hand, speech-related motor cortex modulates the perception of upcoming speech sounds to guarantee fast and fluent production. Neural correlates of human speech perception, production, and their association have been intensively studied in behavior, neuroimaging, and patient experiments. These experimental results suggest an integrative hierarchical framework underlining the sensori-motor processing for speech perception and production, in which the core function is to associate the perceptual and motor representations during speech processing. Based upon this theoretical framework, we built a spiking neural network model including the sensory, motor, and transformation modules, simulating the representations in auditory and motor spaces and their association for processing of American English vowels. The model simulates the online feedback control via an internal model style architecture, the modulation of auditory cortex by motor activation and the internal motor response to perceived speech sounds are realized through the transformation between the representations. By utilizing the spiking neurons as modeling units and applying neuronal level plasticity rules such as spike-timing dependent plasticity and homeostasis, we are able to simulate the learning and dynamics of the association at the neuronal level, and in the long run, bridge the system level cognitive functions with the neuronal activities.