

SPEECH SYNTHESIS

ABSTRACT

Embodiments of the present disclosure provide a solution for speech synthesis. A method comprises: obtaining input information for a trained acoustic autoencoder model, the input information comprising a target text, at least one first reference speech, and at least one second reference speech; extracting, using a timbre encoder, timbre feature information from the at least one first reference speech, the timbre feature information characterizing timbre represented in the at least one first reference speech; extracting, using a prosody encoder, prosody feature information from the at least one second reference speech, the prosody feature information characterizing a prosodic pattern represented in the at least one second reference speech; extracting, using a content encoder, content feature information from the target text; and generating, using a decoder, a target speech corresponding to the target text based on the timbre feature information, the prosody feature information, and the content feature information.

FIG. 4