

SPEECH PROCESSING METHOD, MODEL TRAINING METHOD AND APPARATUSES

ABSTRACT

The present disclosure provides speech processing methods, model training methods, apparatuses,
5 devices, and a storage medium. The method includes: obtaining a to-be-separated speech signal, where
the to-be-separated speech signal is a mixture of a target speech signal of a target speaker and at least
one non-target speech signal of at least one non-target speaker; processing, based on a reference
embedding corresponding to the target speaker, the to-be-separated speech signal with a pre-trained
speech separation model to obtain the target speech signal of the target speaker, where the pre-trained
10 speech separation model includes a separator containing a plurality of processing layers, and the
plurality of processing layers include a first processing layer and remaining processing layers; for a
processing layer of the remaining processing layers, the processing layer is configured to output an
output embedding based on a fused global embedding and an output embedding from a previous
processing layer of the processing layer, and the fused global embedding is refined with the reference
15 embedding corresponding to the target speaker, a fused global embedding from the previous processing
layer and an intermediate separation embedding, where the intermediate separation embedding is
obtained by the processing layer based on the output embedding from the previous processing layer.

FIG. 3