Dear reviewer:

Thanks very much for taking the time to review this manuscript. We really appreciate all your generous comments and suggestions! Here are my detailed answers below.

Point 1. To further improve the paper, the authors should consider more ``selling points" for this work. For example, the whole system consists of separately well-trained two components, but I was wondering if an end-to-end deep learning system, where a deep learning system for language identification is set up on top of another system for speaker separation. In particular, the 1d time-series signal should not be recovered by the decoding operation of speaker separation, and the step of spectrogram features extraction could be faithfully ignored. The authors may refer to the work [1] for the setup of experimental configuration and analyze the system based on the theoretical paper [2].

Response 1:

Thank you very much for your valuable comments. We understand what the reviewer means. In the speech separation task, the last step is to obtain the estimated separated audio by masking the matrix and pointing the input mixed audio feature matrix, after which the separated feature sequence is recovered into an audio file. We can save this step of recovering audio files and directly input the estimated audio spectrogram features into the back-end language recognition network for the recognition task, so that we can construct an end-to-end deep learning model that can be used in real-world scenarios. We conducted experiments in our paper to simulate overlapping multilingual speech scenarios in realistic and complex acoustic scenarios to initially test the feasibility of our idea. And since the size of our existing hybrid dataset is relatively small at the moment, we have not yet implemented an end-to-end system for these two networks. Our next step is to build such an end-to-end deep learning model. We also write about this idea in the future work of our paper.

Point 2. Moreover, the authors may consider the comparison of different loss functions for the speaker separation system. For example, a simple mean absolute error (MAE) or distributional loss. The authors can refer to the papers [3] and [4].

Response 2:

Thanks for your suggestions. And the training objective of our speech separation system is to train to maximize the scale-invariant signal-to-noise ratio (SI-SNR), which is commonly used as a metric to evaluate source separation as an alternative to SDR. SI-SNR is defined as：

$$\begin{cases} s_{target} = \dfrac{\langle \hat{s}, s \rangle s}{\|s\|^2} \\ e_{noise} = \hat{s} - s_{target} \\ SI-SNR = 10\log_{10} \dfrac{\|s_{target}\|^2}{\|e_{noise}\|^2} \end{cases}$$

where $\hat{s} \in R^{1*T}$ and $s \in R^{1*T}$ are the estimated and original clean sources, respectively.

$\|s\|^2 = < s, s >$ denotes the signal power. Scale invariance can be ensured by normalizing$\hat{s}$ and s to zero mean prior to computation.
Thank you for your suggestion and the references provided, I have carefully both of them, our separation system training target is the SI-SNR formula.

Point 3. Additionally, the authors should take care of some minor mistakes in English usage, e.g., As Most of ... (line 12); and experimental setup in Section 3 ... (line 88), etc.

Response 3:

Thanks for your help. We feel really sorry for our carelessness. And we have made changes to the areas pointed out by the reviewer.
After modification:
In line 12, the M in As Most of is changed to a lowercase m;
Section 3 in line 88 repeatedly changed to Section 4.

Yours sincerely