Dear Reviewers:

Thank you for the reviewers’ comments. We have studied comments carefully and have corrected. The responses are as follows:

**Reviewer #1:**

1. As one of the critical contribution claimed by the authors, the quantization related design, including 1) the design space searching for optimal quantization decision; b) hyper-parameter quantization, have been presented in Section 3. However, Section 3 does not provide any specific design solution rather than just pointing to the experimental sections as for the result based analysis. I did not obtain any 'decision making' information from Section 3 for the quantization for the different data for the model.

**Response:**

Thanks for your query for our strategy of searching best quantization decision. The method we follow is based on hardware and accuracy consideration. In order to fit our FPGA platform, the maximum of data word-length should be set as 32 bits. This constraint is the upper bound of our design space research method. On the other hand, experiments show that intermediate results usually need 8~12 bits and normalization parameters need 20~21 bits to indicate their integer part, which is the lower bound of our question. In deep neural networks intermediate results are relatively robust to numerical precision, so we can allocate less mantissa bits for such kind of data. However, normalization process needs more data indication performance than intermediate results, so in principle, we give normalization parameters decimal bitwise no less than intermediate results.

We have supplemented contents above in Section 3.3 in revised version to illustrate our points more clearly. Thanks again for your question!

1. From Fig.3, the the conv-1, conv-2 and FC-1 layers in the accelerator is pipelined for the execution, then how about the rest of the modules (Norm, Vec. Pro. for FC-2, FC-3)?

**Response:**

Thank you for your question on this detail issue and we feel sorry for not introducing our design clearly. In fact, FC-2 and FC-3 function parts do not work in the layer-by-layer pipeline. Conv-1, Conv-2 and FC-1 occupy over 99% of total computing scale, in another word, FC-2 and FC-2 take over a rather little proportion of computing task (the experiment results in Table.2 can also support this idea). In our architecture, to process one audio frame, the time consumed by Conv-1, Conv-2 and FC-1 is over 30k periods while FC-2 and FC-3 only costs hundreds of periods. So there is no need to let FC-2 and FC-3 work in a layer-pipeline mode.

We have modified Fig.3 and added extra explanation about this detail in Section 4.3. Thanks for your subtle question!

1. Another major issue is the style of the Figures and Tables in this manuscript. I hope the authors could carefully refer to other scientific paper for the presentation of the Figures and Tables.

**Response:**

Thanks for your advice. We have remade several figures and tables: Fig.1, Fig.3, Fig.4, Fig.7, Fig.10 and Table.2. Thanks for your helpful suggestion to make our figures more delicate and more clearer.

1. For the quantization results shown in Figure 8, what is the baseline floating point counterpart?

**Response:**

The counterpart is the floating-feature version of target neural network, whose accuracy is 85.31%. We have supplemented this information in the end of Section 3.1. Thanks for your question!

1. The authors may want to notice the format requirement of the journal. It become hard to follow the content when the manuscript has to breakpoint in the middle.

**Response:**

We have changed the relative position between text contents and figures or tables. We follow your suggestions to arrange our contents as fluent as possible. Thanks for your remind!