This study employed a Pretest-Intervention-Posttest design to study the effect of audio-visual training on second language learning, with a focus on pronunciation training. 10 international subjects participated in the study. Based on the preliminary analysis, it is found that (1) The duration of the second syllable of the learners was lengthened significantly with a medium effect size (2) The ratio of Rime2/Rime1 are closer to the native speaker norm after training. (3) Less variance in bi-syllabic word ratio is found after the intervention. (4) Result of two-way ANOVA shows that both token-type (Tone1, Tone3, Tone6) and Mandarin level of participants (Intermediate or above VS beginner) do not moderate the training effect statistically. The result of this study is in line with the plausibility of previous Computer-assisted pronunciation training (CAPT) studies. For Fluency Model, anonymized trials from 29 students in previous elementary courses were extracted for analysis. Overall question completion rate is around 60%. It may suggest that the Fluency Model may need to accommodate slower learners in order to raise the completion rates. In order to identify representational disfluency types among the elementary learners, trials of the students have been transcribed and disfluency types (filler, breakdown, pause) were analyzed. Based on the preliminary analysis, it is found (1) Number of words uttered correlate with breakdown with a medium effect size ($r=0.6$, $p<0.05$) (2) Number of words uttered does not correlate with filler nor pause. Further investigations have to be carried out in order to confirm whether filler is a transfer of first language speaking style. Besides, professional rater judgment and native speaker perception may be used to test whether a relationship exists between disfluency types and listener perception. For Grammar Knowledge Model, further investigations have to be carried out later.