# 先端科学技術研究科　修士論文要旨

| | | | |
|---|---|---|---|
| 所属研究室<br>（主指導教員） | ヒューマンAIインタラクション<br>（Sakriani Sakti （教授）） | | |
| 学籍番号 | 2411801 | 提出日 | 令和 7年 1月 21日 |
| 学生氏名 | 東　翔 | | |
| 論文題目 | Enhancing Switching Dynamics in Transformer-Based Code-Switching ASR | | |

要旨

In recent years, code-switching (CS), where speakers switch languages during conversations, has been frequently observed in multilingual communities and automatic subtitle generation scenarios. Specifically, in cases such as accented word-level CS and non-accented word-level CS, the same pronunciation can have different meanings depending on the language, which has become a major factor in the decline of accuracy in automatic speech recognition (ASR) systems. Additionally, in phrase-level CS involving two languages with distinct phonological systems, the range of CS points expands beyond single words, making the selection of CS points more challenging compared to word-level CS. Moreover, disfluencies and phonetic variations often occur during speech, and the unique characteristics of individual speakers further complicate these challenges. These challenges cause ASR systems to misinterpret contexts, leading to the selection of incorrect words or the failure to resolve lexical ambiguities between languages. While conventional ASR models primarily rely on contextual information to extract language-specific cues, this approach has shown limitations in scenarios involving CS with different accents.

To address these challenges, this study proposes two novel ASR models designed to improve performance in CS scenarios. The first model leverages reinforcement learning (RL). While conventional self-attention mechanisms determine CS points based on context, reinforcement learning optimizes rewards and enables post-hoc learning by using loss as a reward. This property allows the model to adapt flexibly to phonetic variations unique to individual speakers. In this model, language information extracted using a self-attention mechanism is input as a binary sequence, and Q-learning is employed to dynamically select language adapters on a frame-by-frame basis. By designing a simplified action space, the model optimizes recognition accuracy as a reward by selecting the appropriate language adapter for each frame. This is expected to improve the accuracy of multilingual speech recognition involving CS while reducing dependency on context and pronunciation.

The second model utilizes CTC Loss in intermediate layers. Connectionist Temporal Classification (CTC) Loss is a loss function used in speech recognition to align phoneme sequences with acoustic sequences. However, traditional CTC-based models calculate loss only at the final layer, lacking language specificity in CS scenarios. In this study, we propose a novel method that applies CTC Loss to intermediate layers. This enables the detection of token errors in early layers, facilitating language identification. Subsequent layers can then focus on processing the identified language, thereby improving recognition accuracy.

Experimental results showed that the RL model did not demonstrate a significant advantage in CS recognition accuracy compared to the CTC-based model. However, the model utilizing intermediate CTC achieved notable improvements in CS recognition accuracy, demonstrating the effectiveness of this approach in multilingual speech recognition involving CS.