ISS-P-043

ITA、ROHAN4600 コーパスを用いた 日本語の自動読唇精度向上に関する研究

井上　優也†
† 福岡工業大学大学院工学研究科電子情報工学専攻

倪　宝栄††
†† 福岡工業大学工学部電子情報工学科

1. はじめに

今日、英語の自動読唇精度はLipNet[1]の登場により飛躍的に精度が向上した。しかし、日本語の自動読唇においては現状、文章単位での自動読唇の精度は低いと言える。そのため、日本語の自動読唇の精度の向上が必要であると言える。

本研究では機械学習を用いた自動読唇技術及び、ITA、ROHAN4600 マルチモーダルデータベースを用いて文章単位の自動読唇精度向上を目的として行う。

2. 研究手法

2.1. ITA、ROHAN4600 マルチモーダルデータベース

ITA、ROHAN4600 マルチモーダルデータベースとはITA コーパス[2]および ROHAN4600 コーパス[3]を読み上げた動画のフレーム、ランドマーク、音声、音声ラベルデータが含まれており、SSS 合同社(@SSS)が公開している研究目的での使用が可能なマルチモーダルデータベースである。

2.2. データセットの前処理

本研究では上記のデータセットを使用する。使用する内容物としては、フレームおよびランドマークデータ、また ITA、ROHAN4600 コーパスからコーパスデータを用いる。

フレームデータでは口元のみを用いる必要がある。そのため、口元領域の切り出しを Python で行う。また、ランドマークデータも口元のランドマークのみに整形する。

ITA、ROHAN4600 コーパスは形態素解析およびローマ字へ変換を行う。変換後の文字列データからアルファベットと半角スペース以外の文字、記号を削除する。

2.3. 学習方法

学習方法として図 1 のアーキテクチャを用いて学習を行う。パート1及びパート2を用いた学習を行う。パート1に入力としてフレームを、パート 2 には入力としてフレーム及びランドマークを入力する。最終的な出力から損失を算出する際に用いるアノテーションデータとして前処理したコーパスデータを用いる。

3. 結果と考察

学習結果は図 2、図 3 である。

図 2 及び、図 3 からパート 1、パート2共に過学習をしていると言える。原因としてはデータ数の不足が大きいと考える。また、今回用いたアーキテクチャの選定理由については当日発表を行う。

4. 今後の課題

結果と考察からデータの増加方法を検討する。

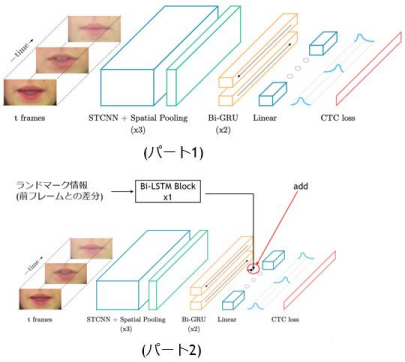

図 1：使用アーキテクチャ

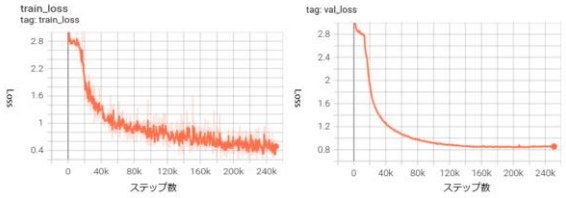

図 2:パート 1 の結果

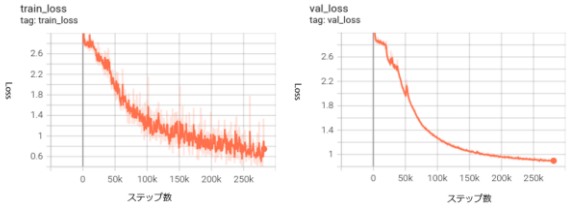

図 3:パート 2 の結果

参考文献

[1] Yannis M. Assael, Brendan Shillingford, ShimonWhiteson, Nando Freitas, “LipNet: End-to-End Sentence-level Lipreading”arXiv:1611.01599, December 2016.

[2] 小口純矢, 金井郁也, 小田恭央, 齊藤剛史, 森勢将雅:ITA コーパス:パブリックドメインの音素バランス文からなる日本語テキストコーパスの構築と基礎評価, 情報処理学会研究報告, vol. 2021-MUS-131, no. 31, pp. 1-6, 2021

[3] 森勢将雅:テキスト音声合成に向けたモーラバランス型コーパスの提案と評価, 日本音響学会 2022 年春季研究発表会, pp. 953-954, Online, March 9-11, 2022