

条件付き確率場を用いた口形遷移情報からの読唇手法の提案と評価

慶應義塾大学 理工学研究科 酒井 悠斗
慶應義塾大学 理工学部 白石 博

概要

本研究は、日本語での会話時における、唇の動き、口形の遷移情報から、実際の会話文の書き起こしを目指したものである。一般にはこの分野は読唇や Lip reading などと呼ばれている。日本語には、母音と子音が存在するが、主に母音により口形が決定している。そのため、口形から母音を推測することは比較的容易であるが、子音の推測は難しいことが知られている。本講演では、口形の遷移情報から子音の系列を推測するモデルを提案する。モデルには、自然言語処理などで用いられることの多い、条件付き確率場を用いる。条件付き確率場によって、母音の系列から子音の系列を推測することで、母音と子音の組み合わせにより、言葉の復元を目指す。

1 はじめに

読唇 (Lip Reading) は、人間の唇の動きから、会話の内容を読み取る技術である。スパイ映画の中の話だと思われるかもしれないが、現に 2016 年には google が LipNet [1] と呼ばれるニューラルネットワークによる読唇の技術を発表した。日本では、[2]などで同様の取り組みが行われているが、英語のような精度は出ていない。今回は、条件付き確率場 (CRF; Conditional Random Fields) というモデルを用いて、この課題に取り組んだ。

2 モデルの説明

一般的に、条件付き確率場、CRF というと Linear Chain CRF という最も基本的なモデルを指すことが多い。これは、ある系列が観測された際に、この観測系列の要素に一对一に対応するような観測できない系列、状態系列が並行して存在すると仮定し、観測系列の情報だけから、状態系列を推測することを目

的とするモデルである。今回のデータに当てはめて説明すれば、「とうけい」という言葉を例にとると、母音の列「o, u, e, i」から子音の列「t(o), -(u), k(e), -(i)」を推測することになる。しかし、このモデルの場合、前後のつながりを考慮はするが、推測の結果が意味のある言葉になっている保証はない。そこで、単語ベースでの解析を行うために、Semi-Markov CRF という発展的なモデルを用いる。このモデルでは単語を 1 まとまりとして認識するため、意味のある単語の列が出力されることが期待できる。

3 データ解析

今回は、会話データとして、名大会話コーパス [3] をもととした。このテキストをカナに変換した後、母音と子音を分けて対のデータを作成する。このデータの一部を用いてモデルを学習させ、残りのデータを使用し、どれほど子音の推測が正しく行えているのかを確認する。詳細は当日に報告する。

参考文献

- [1] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.
- [2] 宮崎剛, 中島豊四郎ほか. 口形ベースの機械読唇における単語認識手法の提案と評価. マルチメディア, 分散協調とモバイルシンポジウム 2014 論文集, Vol. 2014, pp. 896–902, 2014.
- [3] 藤村逸子, 大曾美恵子. 大島 ディヴィッド義和. 会話コーパスの構築によるコミュニケーション研究. 藤村逸子, 滝沢直宏 (編), 言語研究の技法: データの収集と分析, pp. 43–72.