中部電力

共同研究
名古屋大学大学院情報学研究科

特許出願中

注視点情報を用いたマルチモーダル音声認識

発話時の注視点に着目 音声認識性能を向上させる

背景・目的

- 目的指向の作業においては、注視点と音声（言語情報）の関連性が高い。このような相関を利用することにより、音声認識性能を向上させる手法を検討した。
- 深層ニューラルネットワークに基づく「End-to-End音声認識」に注視点情報を統合する方式を検討し、提案手法の有効性を評価した。

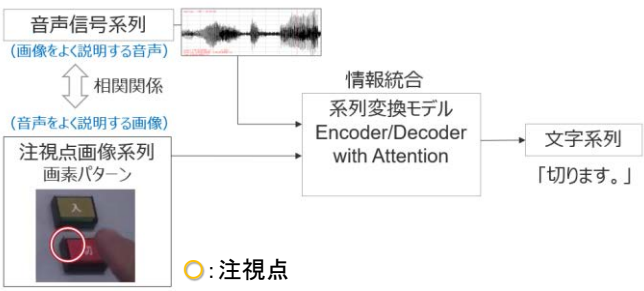
特長

- 系列変換モデルのAttentionの枠組みによって音声情報と注視点情報を統合。
- 話者の主観画像に注視点を表示し、発話内容を同期して関連付けることが可能。

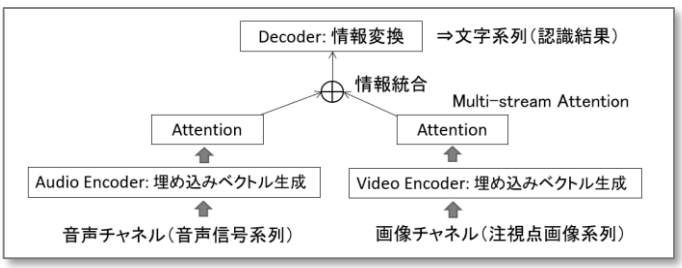
用途

- 音声付き動画へのインデックス付与、シーン検索
- 業務記録（ライフログ）
- 教育研修（行動分析、スキル判定など）

提案手法概要



モデル概要：Attention-based multiple encoder-decoder



評価実験

- 被験者4名による系統模擬操作（発話、注視点）を収録
- 4名のクロスバリデーションによる評価、CER（文字誤り率）の平均
- ベースライン（講演音声で学習）、+話者適応（収集コーパスの1部で転移学習）、+提案手法、の3条件で比較

（操作シーケンス：6発話）

①操作内容の宣言
②操作対象のボタン選択
③選択結果の確認
④操作実行
⑤操作内容の復唱
⑥時間確認

（収集コーパス）学習・評価用データ

話者ＩＤ	セッション数	発話数	文字数
SPK01	23	138	2126
SPK02	20	120	1765
SPK03	20	120	1804
SPK04	20	120	1737
合計	83	498	7432

（評価結果）　文字誤り率

Model	CER [%]
CSJ pretrain （ベースライン）	31.8
+fine-tuning （＋話者適応）	3.9
+video encoder （＋提案手法）	**3.4**

文字誤りの12.8%を削減

開発者のひとこと

- 音声と口唇画像などを併用するアイデアは多数提案されており、最近ではニューラルネットの枠組みによって様々な情報統合の可能性が探求されています。

Copyright © 2018-2020 Chubu Electric Power Co., Inc. and Nagoya Univ. All rights reserved.