Vol.2019-SLP-128 No.9
2019/7/20

発話時の注視点を併用した Attention による音声認識手法の検討

瀬川 修[1]　林 知樹[2]　武田 一哉[2]

概要：本研究では、発話内容と発話時の注視対象との間に何らかの相関関係があると仮定し、音声信号と注視点から成るマルチモーダル情報を統合的に利用する新たな音声認識の枠組みを提案する。本手法では、発話の音響特徴量系列、及び発話区間に対応する注視点の周辺画像の特徴量系列を抽出し、今回提案する Attention-based multiple encoder-decoder ネットワークへと入力する。これにより、音声と注視点という 2 つの異なるモダリティの統合が可能となり、音声認識性能の改善が期待される。提案手法の評価のため、電力系統操作の模擬タスクを設定し、操作中の音声と発話区間に対応する注視点を記録したコーパスを収集した。これを用いた評価実験では文字誤り率の低減が確認され、マルチモーダル情報として注視点を用いることの有効性が示唆される結果が得られた。

キーワード： End-to-End 音声認識、Attention、マルチモーダル、注視点

1. はじめに

本研究では、発話内容と発話時の注視対象との間に何らかの相関関係があると仮定し、音声信号と注視点から成るマルチモーダル情報を統合的に利用する新たな音声認識の枠組みを提案する。何か目的を持った作業等においては、視線と言語情報は相互に関連している可能性が高く、相互の同期関係を推定することにより、情報の補完や予測に有効活用することが考えられる。本稿で述べる手法では、音声信号から音響特徴量系列、そして主観映像中の注視点の周辺画像系列をそれぞれ抽出し、各情報の特徴ベクトルを今回提案する Attention-based multiple encoder-decoder ネットワークへと入力する。これにより、音声と注視点という 2 つの異なるモダリティの統合が可能となり、音声認識性能の改善が期待される。提案手法の評価のため、電力系統操作の模擬タスクを設定し、音声と各発話に対応した注視点の周辺画像からなるコーパスを収集した。以下本稿では、前述のコーパスを用いた実験的評価により、注視点画像系列の併用が音声認識性能に寄与することを検証する。

2. 関連研究

センシング技術の発展と共に、画像や音声、生体信号、そして視線など様々な種類の信号を同時に取得することが容易になりつつある。このような背景の下、音声認識においても音声信号以外のマルチモーダル情報を併用することで認識性能の向上を図る取り組みが数多く提案されている。

2.1 画像情報を利用した研究

代表的なマルチモーダル信号の例として、画像情報が挙げられる。雑音環境下での音声認識性能の改善を目指し、口の動きを表す口唇画像と音声信号を併用する音声認識手法が数多く検討されている [5,6,7]。Mroueh ら[5]は、音響特徴量と口唇画像を別々のニューラルネットワークに入力し、それぞれのネットワークの事後確率を統合することで認識性能の改善を図っている。Noda らの研究 [6]では、Denoising Autoencoder (DAE)を用いて雑音重畳音響特徴量からクリーン音響特徴量を推定し、音素分類モデルとして学習された、口唇画像を入力とする CNN から音素事後確率を計算している。得られたクリーン特徴量と音素事後確率はマルチストリーム隠れマルコフモデルの入力として利用され、GMM-HMM に基づく音声認識システムを構築することで雑音環境下での認識性能の向上を図っている。Petridis ら[7]は、音声信号と口唇画像を単一のネットワークで処理し、直接特定のワードクラスを推定する End-to-End 方式のモデルを提案している。これらの研究では、画像の併用による認識性能の改善が確認されているが、口唇以外の画像を利用することは検討されていない。

2.2 視線情報を利用した研究

その他のマルチモーダル情報の一例として、視線計測デバイスから取得できる「注視点」が挙げられる。例えば、Nguyen らの研究 [8]では、教材ビデオを見ているユーザの注視点から、重要と思われる箇所の文字列を抽出してアノテーションを自動生成し、ユーザが講義ノートを作成する支援機能を提案している。また、視線情報を活用した「映像アノテーション」の事例としては、Vasudevan らが、ユーザ発話(テキスト)、注視点、画像深度、モーション解析等の情報を入力として、ユーザ発話に対応する画像上の領域を同定(Bounding Box を付与)する End-to-End の深層学習の手法 [9]を提案している。この研究では、注視点と言語情報の併用によって、画像上のオブジェクト検出の性能向上を図ることを目的としており、言語情報は所与の入力情報である。これらの研究に対し、注視点と言語情報(音声信号)の併用によって音声認識の性能向上を図るアプローチも考えられるが、これについてはまだ十分な検討がなされていない。

1 中部電力株式会社 エネルギー応用研究所
Chubu Electric Power Co, Inc.
2 名古屋大学 情報学研究科
Graduated School of Informatics, Nagoya University

Vol.2019-SLP-128 No.9
2019/7/20

3. Attention に基づく音声認識

近年、研究開発レベルでは End-to-End 方式の音声認識の検討が盛んに行われている。初期の End-to-End 音声認識では Connectionist Temporal Classification (CTC) [1,2]などに基づく手法が用いられていたが、Attention に基づく手法 [3,4]の有効性が示されるようになり、現在も同手法の発展が続いている。この方式では、入力系列から出力系列へ直接マッピングを行うために、Encoder-Decoder ネットワーク構造が用いられる [11,12]。エンコーダネットワークは、入力特徴量系列を識別的な隠れ状態ベクトル系列へと変換し、デコーダネットワークは Attention 機構を利用して隠れ状態ベクトルと出力系列の要素との間のアライメントを取る。続いて、そのアライメントに基づいて重み付け和された隠れ状態ベクトルを入力として、出力シンボルを推定する。CTC に基づく方式と比較すると、Attention に基づく方式は条件付き独立性の仮定を全く必要としない。さらには、言語モデルや複雑なデコーディング処理も不要となる。この Attention は元々機械翻訳の分野で提案されたモデルであり、入力と出力で語順が入れ替わるような非因果的な状況にも対応可能な非常に柔軟性の高い手法である。しかしながら、音声認識分野では入力と出力の間には必ず因果性が成り立つため、この非因果的なアライメントが問題となる。この問題に対処するため、Watanabe らの手法 [10]では、Attention モデルの目的関数と CTC の目的関数を組み合わせることで、Attention 機構の柔軟なアライメントに対して制約を与えている。

一方 Chiu らの手法 [13]では、より適切なアライメントを得るため Multi-head Attention (MHA)を利用している。MHA では、複数の Attention が計算され、その後単一の Attention に統合される。MHA を利用することで、ニューラルネットワークが異なる時刻における異なる特徴空間上の情報に注目することを可能とし、認識性能が向上するとされている [14]。しかしながら、これらの研究では、新しいネットワーク構造による認識性能の改善が確認されている一方で、マルチモーダル信号の利用は十分に検討されていない。

4. Attention-based multiple encoder-decoder

4.1 提案手法の概要

本研究では、音声特徴量系列と注視点画像系列の 2 つの時系列情報を利用して直接文字系列を推定する Attention-based multiple encoder-decoder に基づくマルチモーダル End-to-End 音声認識の枠組みを提案する。提案手法の概要を図 1 に示す。

提案手法では、音響特徴量系列と注視点画像系列に対してそれぞれエンコーダネットワークを割り当て、それぞれ

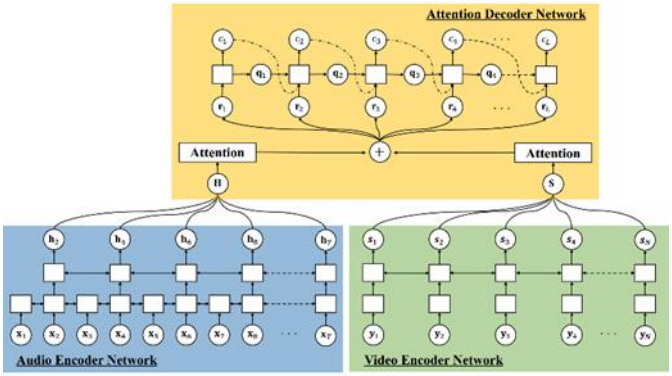

図 1　Attention-based multiple encoder-decoder

の隠れ状態系列に対して Attention 機構を適用する。得られた Attention 重みと隠れ状態系列から文字単位の隠れ特徴量を計算し、デコーダネットワークへ入力することにより認識結果を得る。ここで、音響特徴量はメルフィルタバンクなどの時系列特徴量である。また、注視点画像は注視領域を時系列に並べた画像系列であり、視線計測デバイスで取得された注視点座標を中心に切り出された主観画像内の矩形領域を指す。Attention 機構を音響特徴量系列と画像系列それぞれに対して適用することで、両者の時間分解能の違いを吸収することが可能となる。さらには、両者の発生タイミングのずれの補正が自動的に学習されることも期待される(例えば、物体を注視してから関連する発話を行う場合など)。このように、音声信号と注視点画像系列という 2 つの異なるモダリティを統合した特徴量をデコーダネットワークの入力とすることで、認識性能の改善が期待される。

4.2 定式化

音声特徴量系列 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_T\}$、及び画像系列 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\dots,\boldsymbol{y}_N\}$ が与えられたときに、文字系列 $\boldsymbol{C}=\{c_1,c_2,\dots,c_L\}$を出力する事後確率$p(\boldsymbol{C}|\boldsymbol{X},\boldsymbol{Y})$は確率の連鎖律を用いて次のように分解できる。

$$p(\boldsymbol{C}|\boldsymbol{X},\boldsymbol{Y})=\prod_{l=1}^{L}p\left(c_l|c_{1:l-1},\boldsymbol{X},\boldsymbol{Y}\right) \tag{1}$$

ここで、$c_{1:l-1}$は部分系列 $\{c_1,c_2,\dots,c_{l-1}\}$を表し、$p(c_l|c_{1:l-1},\boldsymbol{X},\boldsymbol{Y})$ は次のように計算される。

$$\boldsymbol{H}=\text{AudioEncoder}(\boldsymbol{X}) \tag{2}$$

$$\boldsymbol{S}=\text{VideoEncoder}(\boldsymbol{Y}) \tag{3}$$

$$a_{lt}=\text{LocationAttention}(\boldsymbol{q}_{l-1},\boldsymbol{h}_t,\boldsymbol{a}_{l-1}) \tag{4}$$

$$b_{ln}=\text{LocationAttention}(\boldsymbol{q}_{l-1},\boldsymbol{s}_n,\boldsymbol{b}_{l-1}) \tag{5}$$

$$\bar{\boldsymbol{h}}_l=\Sigma_{t=1}^{T}a_{lt}\boldsymbol{h}_t \tag{6}$$

$$\bar{\boldsymbol{s}}_l=\Sigma_{n=1}^{N}b_{ln}\boldsymbol{s}_n \tag{7}$$

$$\boldsymbol{g}_l=\sigma(\boldsymbol{W}_g[\bar{\boldsymbol{h}}_l^{\top},\bar{\boldsymbol{s}}_l^{\top}]^{\top}+\boldsymbol{b}_g) \tag{8}$$

$$\boldsymbol{r}_l = \bar{\boldsymbol{h}}_l + \boldsymbol{g}_l \odot \bar{\boldsymbol{s}}_l \tag{9}$$

$$p(c_l|c_{1:l-1}, \boldsymbol{X}, \boldsymbol{Y}) = \mathrm{Decoder}(\boldsymbol{r}_l, \boldsymbol{q}_{l-1}, c_{l-1}) \tag{10}$$

ここで、式(2)及び式(3)はそれぞれ音響エンコーダネットワーク及び画像エンコーダネットワークを表し、式(10)はデコーダネットワークを表す。$\boldsymbol{h}_t, \boldsymbol{s}_n$及び$\boldsymbol{q}_l$はそれぞれ音響エンコーダネットワークの隠れ状態ベクトル、画像エンコーダネットワークの隠れ状態ベクトル、そしてデコーダネットワークの隠れ状態ベクトルを表し、$\boldsymbol{H}$及び$\boldsymbol{S}$はそれぞれ隠れ状態ベクトル系列$\{\boldsymbol{h}_1, \boldsymbol{h}_2, \dots, \boldsymbol{h}_T\}$及び$\{\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_N\}$を表す。$a_{lt}$及び$b_{ln}$はそれぞれ音響エンコーダネットワークの隠れ状態ベクトル系列に対する Attention 重み、画像エンコーダネットワークの隠れ状態ベクトル系列に対する Attention 重みを表す。$\bar{\boldsymbol{h}}_l$と$\bar{\boldsymbol{s}}_l$は、Attention 重みにより重み付け和された文字単位の音響エンコーダ及び画像エンコーダネットワークの隠れ状態ベクトルである。$\boldsymbol{g}_l$は画像エンコーダネットワークの隠れベクトルを利用するかどうかを決定するゲートの役割をもつベクトルであり、最終的に文字単位の音響エンコーダの隠れ状態$\bar{\boldsymbol{h}}_l$とゲート$\boldsymbol{g}_l$で重み付けされた文字単位の画像エンコーダの隠れ状態$\bar{\boldsymbol{s}}_l$を足し合わせることで、デコーダへの入力ベクトル$\boldsymbol{r}_l$が計算される。

音響エンコーダネットワークは、音響特徴量系列$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_T\}$を識別的な隠れ状態ベクトル系列$\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \dots, \boldsymbol{h}_T\}$へと変換するニューラルネットワークであり、畳み込みニューラルネットワークの一種である VGG [15]と Bidirectional Long Short-Term Memory (BLSTM)によってモデル化される。ここで VGG は構造を簡略化した畳み込み層 4、プーリング層 2 のネットワークを用いた。

$$\mathrm{AudioEncoder}(\boldsymbol{X}) = \mathrm{BLSTM}(\mathrm{VGG}(\boldsymbol{X})) \tag{11}$$

音声認識の場合、入力系列の長さが出力系列の長さと大きく異なるため、VGG の max-pooling により入力系列の長さを 4 分の 1 にする。

画像エンコーダネットワークは、画像系列$\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N\}$を識別的な隠れ状態ベクトル系列$\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_N\}$へと変換するニューラルネットワークであり、AlexNet [16] と BLSTM を組み合わせたネットワークでモデル化される。

$$\mathrm{VideoEncoder}(\boldsymbol{Y}) = \mathrm{BLSTM}(\mathrm{AlexNet}(\boldsymbol{Y})) \tag{12}$$

画像の学習データ量の不足に対処するため、AlexNet は予め大規模画像データセットの一つである ImageNet [17]で事前学習したものを用いる。

Attention 重みa_{lt}は、出力系列の要素c_lと音響エンコーダネットワークの隠れ状態ベクトル$\boldsymbol{h}_t$の間のアライメントを表す。$\mathrm{LocationAttention}(\cdot)$は Location-based Attention [3]であり、次のように計算される。

$$\boldsymbol{F}_l = \boldsymbol{K} * \boldsymbol{a}_{l-1} \tag{13}$$

$$e_{lt} = \boldsymbol{g}^{\mathrm{T}} \tanh(\boldsymbol{W}_q \boldsymbol{q}_l + \boldsymbol{W}_h \boldsymbol{h}_t + \boldsymbol{W}_f \boldsymbol{f}_{lt} + \boldsymbol{b}) \tag{14}$$

$$\boldsymbol{a}_l = \mathrm{Softmax}(\boldsymbol{e}_l) \tag{15}$$

ここで$\boldsymbol{F}_l$はベクトル系列$\{\boldsymbol{f}_{l1}, \boldsymbol{f}_{l2}, \dots, \boldsymbol{f}_{lT}\}$を表し、$\boldsymbol{K}$は学習可能な畳み込みフィルタを表す。Attention 重みb_{ln}は、出力系列の要素c_lと画像エンコーダネットワークの隠れ状態ベクトル$\boldsymbol{s}_n$の間のアライメントを表し、a_{lt}と全く同様の手順で計算される。

デコーダネットワークは、RNN 言語モデル [18]と同様に、一つ前の文字c_{l-1}と自身の隠れ状態ベクトル$\boldsymbol{q}_{l-1}$、そして、文字単位の隠れ状態ベクトル$\boldsymbol{r}_l$から次の文字c_lを推定する。デコーダネットワークは LSTM によってモデル化される。

$$\boldsymbol{q}_l = \mathrm{LSTM}(c_l, \boldsymbol{q}_{l-1}, \boldsymbol{r}_l) \tag{16}$$

$$\mathrm{Decoder}(\cdot) = \mathrm{Softmax}(\boldsymbol{W}\boldsymbol{q}_l + \boldsymbol{b}) \tag{17}$$

ここで$\boldsymbol{W}$と$\boldsymbol{b}$はそれぞれ学習可能な行列及びベクトルパラメータを表す。

最終的に、ネットワーク全体は Back-Propagation Through Time (BPTT) [19]を用いて次の目的関数を最小化するように学習される。

$$\begin{aligned}\mathcal{L} &= -\log p(\boldsymbol{C}|\boldsymbol{X}, \boldsymbol{Y}) \\ &= -\log \left(\sum_{l=1}^{L} p(c_l|c^*_{1:l-1}, \boldsymbol{X}, \boldsymbol{Y}) \right)\end{aligned} \tag{18}$$

ここで$c^*_{1:l-1} = \{c^*_1, c^*_2, \dots, c^*_{l-1}\}$は前の時刻の正解文字系列を表す。

5. 評価実験

5.1 コーパス収集

提案手法の有効性評価のため、グラス型の視線計測デバイス Tobii Glass2 (https://www.tobiipro.com/ja/)を用い、電力系統操作の模擬タスクにより操作中の音声に対応する注視点画像を収集した（内蔵カメラによる主観画像中の注視点座標を中心に 128×128pixel の領域を切り出した）。Tobii Glass2 では、注視点 (gaze point)のサンプリングレートは 50Hz で、主観映像(1920×1080 pixel, MP4)のフレームレートは 25fps である。このため、注視点のサンプルを一つ置きに間引いて使用した。

今回設定した「系統模擬操作」のタスクを以下で説明する。本タスクでは被験者が給電制御所のコンソールの系統操作を模擬した一連の手順を行う。当該タスクでは、操作と同時に作業者が自ら行う操作内容や操作結果を発声し、操作結果を確認するという図 3 のシーケンスを繰り返す。実験用に用意した模擬操作パネル外観と視線計測デバイスによる注視点の検出例を図 2 に示す。図中の〇の中心が注視点の座標である。また、発話区間に対応する注視点画像系列の例を図 4 に示す。

図 2 模擬操作パネルと注視点の検出例

① 操作内容の宣言
「猪高で猪高猪子石 1 号線ラインスイッチ 782 を入れます。」
② 操作対象のボタン選択
「ラインスイッチ 782 選択。」⇒ボタン押下
③ 選択結果の確認
「ラインスイッチ 782 操作選択よし。」
④ 操作実行
「入れます。」⇒ボタン押下
⑤ 操作内容の復唱
「猪高で猪高猪子石 1 号線ラインスイッチ 782 を入れました。」
⑥ 時間確認
「時間 15 時 35 分。」

図 3 操作シーケンスと発話例

発話:「遮断器 7 3 4 選択」

発話:「時間 18 時 49 分」

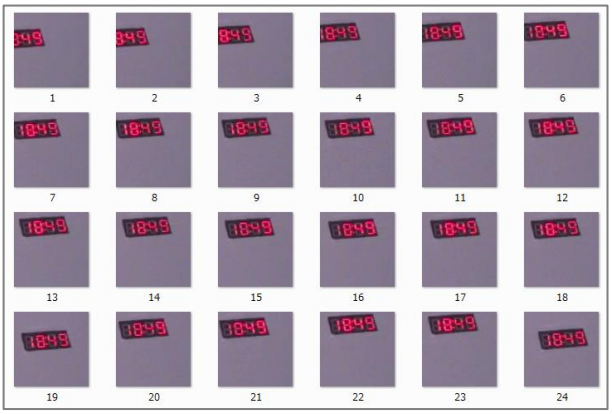

図 4 発話区間に対応する注視点画像系列の例

表 1 コーパス概要

Speaker ID	Num. of Session	Num. of Utterance
SPK01	23	138
SPK02	20	120
SPK03	20	120
SPK04	20	120

表 1 に収集したコーパスの概要を示す。コーパス収集では、前述の図 3 のシーケンスを 1 セッション(6 発話)とし、被験者ごとに複数セッションのデータを収録した。被験者は男性 4 名(当該分野の業務経験なし)である。なお、収集ではセッションごとに線路名と各スイッチの数字を貼り替えて収録を行った。

5.2 実験条件

前述のコーパスを用いて、leave-one-subject-out 検証に基づく実験的評価を実施した。4 名の被験者のうち、2 名を学習データとし、残り 2 名を、それぞれ検証データ及び評価データとして用いた。評価データに用いる被験者を順番に入れ替えて合計で 4 回の評価を行い、その平均を最終的な評価値として用いた。

実験条件の詳細を表 2 に示す。今回の検討では、文字単位の Hybrid CTC/Attention architecture [10]の実装系 (ESPnet [22]) をベースに、提案手法の Attention-based multiple encoder-decoder を実装した。学習データの量が比較的少量であることから、日本語話し言葉コーパス(Corpus of Spontaneous Japanese: CSJ) [20]を用いた事前学習を実施した。まず、CSJ の全講演データを用いて、画像エンコーダ部分を除いたモデルの学習を行った。その後、学習した音響エンコーダ部分とデコーダ部分の重みを初期値として利用し、画像エンコーダ部分を含めたモデル全体の再学習を行った。なお、画像エンコーダの CNN による特徴量抽出部は、ImageNet [17]を用いて学習された重みを初期値として利用した。また、音響エンコーダの学習には、CTC の目的関数を合わせて利用する Joint CTC-Attention multi-task learning を利用した [10]。評価尺度には、式(19)で表される character error rate (CER)を用いた。

$$\mathrm{CER} = \frac{S + D + I}{N} \tag{19}$$

ここで、S,D,I 及び N はそれぞれ置換誤り数、削除誤り数、挿入誤り数、そして全文字数を表す。

提案手法の有効性を確認するため、以下の 2 種類のモデル（画像エンコーダ無し）との比較を行った。

1. CSJ の全講演データを用いて学習したモデル
2. CSJ の全講演データを用いて学習した重みを初期値として、評価タスクのデータで再学習（fine-tuning）したモデル

表 2 実験条件

# CSJ training data	445,068
# utterances in our database	498
# unique characters	3,260
sampling rate	16,000 Hz
window size	25 ms
shift size	10 ms
acoustic encoder type	VGG -BLSTM
# acoustic encoder BLSTM layers	4
# acoustic encoder BLSTM units	2,048
# acoustic encoder projection units	1,024
video encoder type	AlexNet-BLSTM
# video encoder BLSTM layers	1
# video encoder BLSTM units	2,048
# video encoder projection units	1,024
acoustic encoder Attention type	Location-based
kernel size in acoustic encoder Attention	100
# filters in acoustic encoder Attention	10
image encoder Attention type	Location-based
kernel size in video encoder Attention	20
# filters in video encoder Attention	10
decoder type	LSTM
# decoder layers	1
# decoder units	1,024
learning rate	1.0
dropout	0.2
gradient clipping norm	5
batch size (pretrain)	20
batch size (finetune)	8
maximum epoch	15
optimization method	AdaDelta [21]
AdaDelta ρ	10^{-8}
AdaDelta ϵ	10^{-2}
beam size	20
MTL alpha	0.5
CTC weight in decoding	0.3

表 3 実験結果

Model	S %	D %	I %	CER %
CSJ pretrain	23.6	2.5	7.0	33.0
+fine-tuning	5.3	0.9	1.0	7.2
+video encoder	5.2	1.1	0.7	**6.9**

5.3 実験結果

実験結果を表 3 に示す。CSJ のみで学習を行ったモデル(CSJ pretrain)の性能が非常に低いが、これは主に評価タスク固有の未知語の影響が大きいためである。評価タスクのコーパスを用いた再学習(+fine-tuning)による大幅な性能改善は、音響的な適応効果に加え、この未知語の影響が低減されたことに起因すると考えられる。

提案手法の効果であるが、表 3 の結果(+video encoder)より画像エンコーダを用いることによって CER が 7.2%から 6.9%に改善された。これにより、Attention に基づく音声認識にマルチモーダル情報として注視点を用いることの有効性が示唆される結果が得られた。

6. おわりに

本研究では、発話内容と発話時の注視対象との間に何らかの相関関係があると仮定し、音声信号と注視点からなるマルチモーダル情報を統合的に利用する新たな音声認識の枠組みを提案した。また、提案手法の有効性評価のため、電力系統操作の模擬タスクを設定し、操作時の音声と主観映像中の注視点の周辺画像から成るコーパスを収集した。提案手法と前記コーパスを用いた評価実験では、Attention ベースの音声認識に注視点情報を統合する方式の有効性が示唆される結果が得られた。今後の課題としては、注視点画像系列の Attention 重みの分析、及び大規模コーパスを用いた評価などが挙げられる。

参考文献

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*.ACM, 2006, pp. 369–376.

[2] A.Graves and N.Jaitly, “Towards End-to-End speech recognition with recurrent neural networks,” Proceedings of the 31st International Conference on Machine Learning, 2014, pp.1764-1772.

[3] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, “End-to-End continuous speech recognition using attention-based recurrent NN: First results,” *arXiv preprint arXiv:1412.1602*, 2014.

[4] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in *Advances in neural information processing systems*, 2015, pp. 577–585.

[5] Y. Mroueh, E. Marcheret, and V. Goel, “Deep multimodal learning for audio-visual speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE*

International Conference on.IEEE, 2015, pp. 2130–2134.

[6] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[7] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-End audiovisual speech recognition," *arXiv preprint arXiv:1802.06424*, 2018.

[8] C. Nguyen and F. Liu, "Gaze-based notetaking for learning from lecture videos," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.ACM, 2016, pp. 2093–2097.

[9] A. B. Vasudevan, D. Dai, and L. Van Gool, "Object referring in videos with language and human gaze," *arXiv preprint arXiv:1801.01582*, 2018.

[10] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention architecture for End-to-End speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[11] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[13] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," *arXiv preprint arXiv:1712.01769*, 2017.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009.

[18] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[19] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[20] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese." in *LREC*.Citeseer, 2000.

[21] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[22] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "ESPnet: End-to-End speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.