# マルチモーダル音声認識における Local attention を用いた音声画像統合方式*

☆角田遼太 (神戸大), 相原龍 (三菱電機), 高島遼一,
滝口哲也 (神戸大), 本山信明 (三菱電機)

## 1 はじめに

近年，深層学習技術の発展によりニューラルネットワークを用いた音声認識の精度が大幅に向上している．音声認識はスマートフォンにおける検索機能や，カーナビゲーションの音声操作などに利用されている．しかし，音声信号が劣化するような雑音環境下では，発話内容の認識精度が大きく低下するという課題がある．

一方，人間は発話内容を認識する際に様々な情報を統合的に利用している．特に唇の情報の影響は大きく，例えば人間は唇の動きと音声が一致しない映像を見た場合，発話内容を誤って理解してしまう McGurk effect(マガーク効果) が知られている [1]．そのため人間は音声が聞き取りにくい場合であっても，唇の情報から発話内容をある程度理解できる．このことから，雑音環境下での発話内容の理解には，音声と口唇動画像の統合的利用が有用と考えられる．

上記のような背景から，音声と口唇動画像を併用することで，発話認識の精度向上を目的とする，マルチモーダル (Audio-Visual) 音声認識の研究がされている．マルチモーダル音声認識は，特に雑音環境下においてその有効性が示されている [2]．そのため，車載カメラの映像を併用した音声認識など，雑音が大きいと想定される環境に適用することが期待される．

マルチモーダル音声認識において，音声と口唇動画像の情報を統合する様々な手法が提案されている．文献 [3] では，音声と口唇動画像から得られた特徴量を用いて Hybrid CTC/attention モデルで発話内容を認識する手法を提案し，モデルの内部で統合する手法 (early fusion) と外部で統合する手法 (late fusion) の比較を行っている．また，文献 [4, 5] では，音響特徴量をクエリとして画像特徴量に Attention をかける Cross-modal attention 機構により音声と画像の特徴量を統合し，Encoder-Decoder モデルで音声認識を行う AV Align が提案されている．一般に，動画像のフレームレートは音声のサンプリングレートに比べて低いため，early fusion では画像のアップサンプリングが必要であるが，Cross-modal attention 機構を使用することで，アップサンプリングを行うことなく音声と画像の特徴量を統合することができる．本研究では AV Align を Baseline とする．

従来の AV Align では，Cross-modal attention を計算する際に，全ての時刻の画像フレームに対して重み付けを行っている．この手法を Global attention と呼ぶことにする．一般に Attention は雑音の大きい音声では重みの推定が困難であることが知られているため [6]，Global attention では雑音環境下において正確な重みの推定が困難な可能性がある．そこで本研究では，ある時刻の音声に関係する画像のフレームは全体の一部であるという考えに基づき，重みの計算を一部の画像フレームに限定する Local attention [7] を使用することを提案する．Local attention であれば雑音環境下であっても重み計算を行うフレームを限定することで，適切な重みの推定が可能であることが期待される．

以下，第 2 章で Cross-modal attention 機構について紹介する．第 3 章で Local attention を用いた音声・画像の統合方法について述べ，第 4 章で評価実験を行い，第 5 章で本稿をまとめる．

## 2 Cross-modal attention 機構

一般に，音声と動画像はフレームレートが異なるため，early fusion を行うためには音声・画像間のフレームの対応（アライメント）を得る必要がある．Cross-modal attention 機構は文献 [4] で提案されている，音声・画像間のアライメントを自動的に学習する手法である．まず初めに，音響特徴量 $\boldsymbol{a} = \{a_1, a_2, \ldots, a_N\}$，画像特徴量 $\boldsymbol{v} = \{v_1, v_2, \ldots, v_M\}$ がそれぞれ Audio

* Audio-image integration using local attention for audio-visual speech recognition. by Ryota Tsunoda (Kobe University), Ryo Aihara (Mitsubishi Electric Corporation), Ryoichi Takashima, Tetsuya Takiguchi (Kobe University), Nobuaki Motoyama (Mitsubishi Electric Corporation)

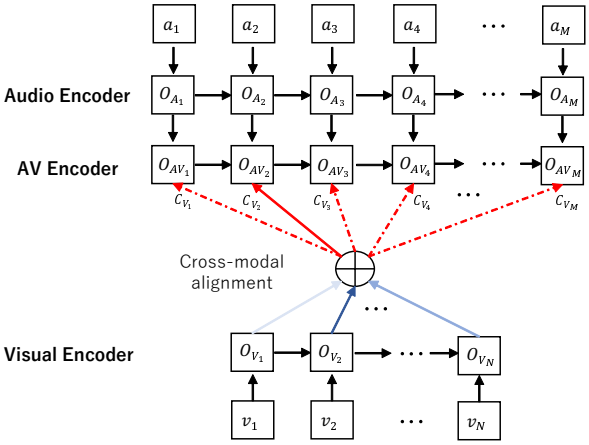


Fig. 1　Cross-modal attention mechanism

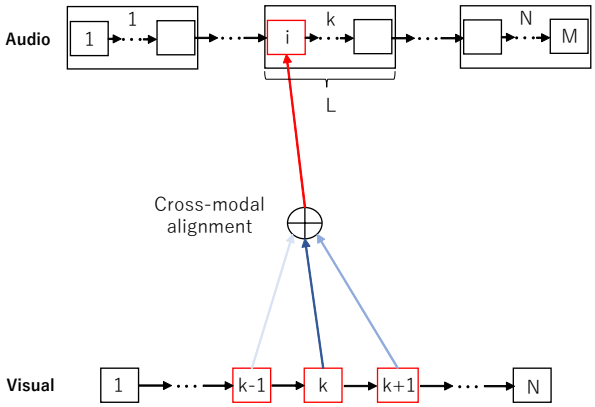


Fig. 2　Proposed method (Window size=3)

Encoder，Visual Encoder へ入力される．それぞれの特徴量は Encoder によって処理され，高次の特徴量 $\boldsymbol{o}_A = \{o_{A_1}, o_{A_2}, \ldots, o_{A_N}\}$, $\boldsymbol{o}_V = \{o_{V_1}, o_{V_2}, \ldots, o_{V_M}\}$ へ変換される．

$$o_{A_i} = \mathrm{Encoder_A}(a_i, o_{A_{i-1}}) \quad (1)$$

$$o_{V_j} = \mathrm{Encoder_V}(v_j, o_{V_{j-1}}) \quad (2)$$

その後，音響特徴量をクエリとして画像特徴量に対して Attention をかけることによって各時刻の音響特徴量に対して関連度の高い画像特徴量を抽出し，コンテキストベクトル $\boldsymbol{c}_V$ を計算する．

$$h_i = \mathrm{Encoder_{AV}}([o_{A_i}; o_{AV_{i-1}}], h_{i-1}) \quad (3)$$

$$c_{V_i} = \mathrm{attention}(h_i, o_V) \quad (4)$$

最後に，コンテキストベクトルと音声の特徴量を統合する．

$$o_{AV_i} = W_{AV}[h_i; c_{V_i}] + b_{AV} \quad (5)$$

このモデルの概要を Fig. 1 に示す

## 3　提案手法

本研究では，Cross-modal attention 機構にて雑音環境下に頑健な音声と画像のアライメントを学習するために，Local attention を使用する手法を提案する．提案手法の概略図を Fig. 2 に示す．一般的に，音声のフレーム数は画像のフレーム数よりも大きい．

まず初めに，音声と画像のフレームレートの違いから，音声のフレームを画像のフレーム数分のグループに分割する．音声のフレーム数を $M$，画像のフレーム数を $N$ とすると $(M \geqq N)$，画像 1 フレームあたり，音声は $L = M/N$ フレームが対応することになる．本手法では音声のフレーム系列を $L$ フレーム毎にグループ化し，グループ毎に Local attention を計算する際のフレーム範囲を定義する．すなわち，時刻 $i$ の音声フレームが割り当てられるグループのインデクス $k$ は以下の式で計算され，

$$k = \lceil \frac{i}{L} \rceil \quad (6)$$

このグループの音声フレームは，$k$ 番目周辺の画像フレームと関係が強いと仮定する．

時刻 $i$ の音声フレームについて Attention を計算する際は，その音声フレームが属しているグループに対応する画像のフレーム $k$ に焦点を定める．この時，$k$ を中心として窓幅 $D$ の大きさ分の画像フレームを使用して Attention を計算し，音声の特徴量と統合する．したがって，Attention を計算する際に 2 章で説明した (4) 式において，

$$c_{V_i} = \mathrm{attention}(h_i, o_{V_{k-\frac{D}{2}:k+\frac{D}{2}}}) \quad (7)$$

とすることで，Local attention を使用して音声と画像の特徴量を統合することができる．Fig. 2 は窓幅が 3 である場合の例を示している．

## 4　評価実験

### 4.1　実験条件

提案手法の有効性を示すためにマルチモーダル音声認識用のデータセットとして TCD-TIMIT [8] を用いた．TCD-TIMIT は 62 人の話者が合計 6913 文を発話している音声とビデオ映像で構成されている．本実験では，TCD-TIMIT

の Speaker-dependent の設定に従い訓練データ数 3752 文，評価データ数 1736 文を用いて学習及び評価を行った．

マルチモーダル音声認識には End-to-End 音声認識ツールキットである ESPnet [9] を用いて，Fig. 3 に示すような Hybrid CTC/Attention モデル [10] の学習を行った．音声の入力特徴量として，23 次元のメルフィルタバンク特徴量にピッチ特徴を合わせた計 26 次元の特徴量，画像の特徴量には，ビデオ映像から OpenFace [11] を用いて顔画像を検出した後，唇領域を切り取って 36×36 にリサイズした 3 チャネルのカラー画像を使用した．出力次元数は、英文字 26 種類にアポストロフィ，未知文字，空白，開始記号および終端記号加えた 31 次元とした．Audio Encoder は 320 次元の隠れ層を持つ 5 層の双方向 GRU，Visual Encoder は文献 [5] で使用されている 11 層の Resnet CNN に続けて，320 次元の隠れ層を持つ 1 層の単方向 LSTM を使用した．そして，AV Encoder には 320 次元の隠れ層を持つ 1 層の単方向 LSTM を使用した．デコーダは 320 次元の隠れ層を 1 つ持つ 1 層の単方向 LSTM と，その後の 31 次元のノードを持つ softmax 層から構成される．Attention 機構には Coverage mechanism location aware attention を使用し，AdaDelta を用いて最適化を行った．マルチタスク学習時には CTC の損失関数の重みを 0.5，認識時の CTC の出力確率の重みも同様に 0.5 に設定した．

本研究では，背景雑音環境下と妨害発話環境下の 2 つの環境下で提案手法の有効性を評価した．背景雑音環境では信号対雑音比を 10 dB，0 dB として TCD-TIMIT の音声データに雑音を重畳した．モデルの学習は clean，10 dB，0 dB の順に行い，雑音環境下では前段階の学習済みモデルの重みを初期値としてファインチューニングを行った．雑音には文献 [5] で使用されている Cafeteria noise を重畳した．妨害発話環境では，TCD-TIMIT から妨害話者を 1 名選択し，目的話者音声に重畳した．なお，口唇動画像は目的話者に対応したもののみが入力される．学習手順は，雑音音声の重畳時と同様である．

### 4.2 実験結果

Table 1 に Cafeteria noise 環境下における，従来手法及び提案手法の文字誤り率を示す．窓幅は 5〜31 に変化させた．口唇動画像を使用することで，音声のみを用いる場合よりも認識精度が改善していることがわかる．信号対雑音比が小さい環境下では，口唇動画像と音声を統合させることで，認識精度が大きく改善されていることが分かる．

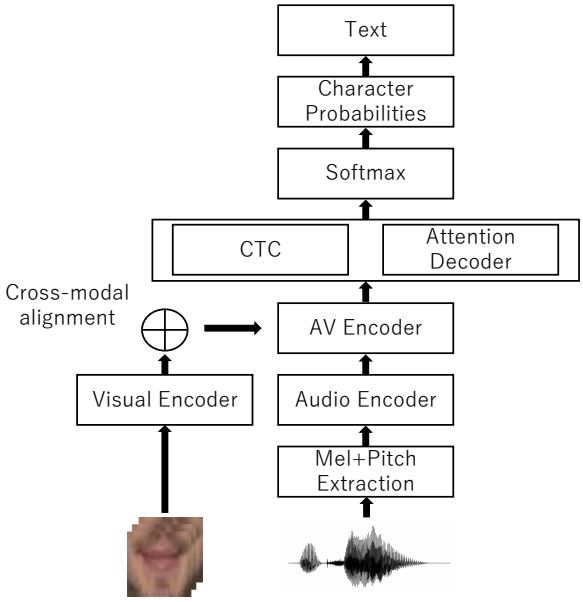


Fig. 3　Model architecture

Table 1　Character Error Rate (CER) in background noise

| Model | Window size | CER[%] clean | 10 dB | 0 dB |
|---|---|---|---|---|
| audio only | - | 24.8 | 35.7 | 52.8 |
| baseline(global) | ∞ | 22.0 | 31.3 | 46.6 |
| local | 5 | 21.8 | 30.3 | 44.1 |
| | 11 | 21.2 | 29.2 | 42.7 |
| | 21 | 21.1 | 29.5 | 43.2 |
| | 31 | 21.7 | 29.2 | 43.8 |

また，提案手法と従来手法を比較すると，Clean 環境では認識精度に大きな差は生じなかった．その一方で，信号対雑音比が小さい環境下に注目すると，従来手法と比較して，相対比で 10 dB では最大 6.7%，0 dB では 8.4%の改善が得られた．このことから，提案手法は特に雑音環境下において有効であると言える．また，Local Attention の窓幅は 11 が最も精度が高かった．

Table 2 に妨害発話環境下での評価結果を示す．Cafeteria noise を重畳している Table 1 の結果と比較して，全体の認識精度が大きく低下していることがわかる．しかしその一方で，従来手法と提案手法を比較した場合，相対比で最大 9.5%の改善が得られたことから，このような難易度の

Table 2　Character Error Rate (CER) with interference speaker

| Model | Window size | CER[%] | |
|---|---|---|---|
| | | clean | 2spk |
| audio only | - | 24.8 | 74.0 |
| baseline(global) | $\infty$ | 22.0 | 65.2 |
| local | 5 | 21.8 | 59.0 |
| | 11 | 21.2 | 60.1 |
| | 21 | 21.1 | 59.2 |
| | 31 | 21.7 | 59.8 |

高いタスクにおいても Local attention を使用する効果があると考えられる．

## 5　おわりに

本研究では，Local attention を用いた音声と画像の統合方法を提案した．提案手法は従来手法と比較して，雑音を重畳していないデータでは大きく差は生じなかったが，背景雑音が存在する環境下では従来手法よりも優位な結果を示した．また，妨害発話が存在する比較的認識難易度の高いタスクにおいても，背景雑音下と比較して認識精度は大きく低下しているものの，提案手法が従来の Global attention を上回る認識精度を示した．

今後の課題として，2 人以上の話者が同時に発話を行っている音声データから，目的話者の発話内容を高い精度で認識するためのモデルを検討することがあげられる．

## 参考文献


[1] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” Nature, Vol. 264, pp. 746–748, 1976.

[2] K. Paleček and J. Chaloupka, “Audio-visual speech recognition in noisy audio environments,” in *Proc. International Conference on Telecommunications and Signal Processing (TSP)*, pp. 484–487, 2013.

[3] S. Petridis *et al*, “Audio-visual speech recognition with a hybrid ctc/attention architecture,” in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pp. 513–520, 2018.

[4] G. Sterpu *et al*, “Attention-based audio-visual fusion for robust automatic speech recognition,” in *Proc. of the 20th ACM International Conference on Multimodal Interaction*, pp. 111–115, 2018.

[5] G. Sterpu *et al*, “How to teach dnns to pay attention to the visual modality in speech recognition,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, Vol. 28, pp. 1052–1064, 2020.

[6] S. Kim *et al*, “Joint CTC-attention based end-to-end speech recognition using multitask learning,” in *Proc. ICASSP*, pp. 4835–4839, 2017.

[7] T. Luong *et al*, “Effective approaches to attention-based neural machine translation,” in *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015.

[8] N. Harte and E. Gillen, “TCD-TIMIT: An audio-visual corpus of continuous speech,” *IEEE Transactions on Multimedia*, Vol. 17, pp. 603–615, 2015.

[9] S. Watanabe *et al*, “ESPnet: End-to-end speech processing toolkit,” in *Proc. Interspeech*, pp. 2207–2211, 2018.

[10] S. Watanabe *et al*, “Hybrid ctc/attention architecture for end-to-end speech recognition,” *IEEE Journal on Selected Topics in Signal Processing*, Vol. 11, No. 8, pp. 1240–1253, 2017.

[11] T. Baltrušaitis *et al*, “Openface 2.0: Facial behavior analysis toolkit,” in *Proc. 13th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 59–66, 2018.