妨害発話環境下におけるマルチモーダル音声認識モデルの事前学習方式の検討 *

☆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), △今井良枝 (三菱電機)

1 はじめに

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．特に発話者の唇の動きから得られる情報が与える影響は大きい．例えば，人間は唇の動きと音声が一致しない映像をみた時，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，唇の動き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，発話内容の理解には，音声と唇の動きの統合的利用が有用と考えられる．

上記の背景から，音声と口唇動画像を併用することで，発話内容の認識精度向上を目的とするマルチモーダル (Audio-Visual; AV) 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雑音の影響で音声信号が劣化する状況において，音声認識の頑健性を向上させることが知られている [2] ため，カーナビゲーションシステム等での利用が期待されている．

従来の研究では，背景雑音環境下における認識性能の改善のために，音声と画像の特徴量を用いて Hybrid CTC/attention モデルで認識を行う手法 [3] や，Attention 機構を用いて音声と画像の特徴量を統合する AV Align[4, 5] 等が提案されてきた．その一方で，複数の話者が同時に発話を行っている妨害発話環境下では，目的話者と他の話者の発話を分離することが困難であるため，背景雑音環境下における認識と比較して認識精度が低くなる．我々が以前に提案した手法 [6] では，AV Align に対して妨害話者の発話内容に基づく補助損失関数を導入することで，目的話者の音声認識精度が向上することを示した．しかし，この手法では目的話者の音声認識と同様に妨害話者の音声認識を行うことができるが，目的話者と比較して妨害話者の音声認識精度は低いため，音源分離の性能が十分でないという課題が残されている．

妨害発話環境下における音源分離の手法として，目的話者の音声サンプルと口唇動画像を補助情報とすることで混合音声から目的の音声信号を抽出する Multimodal Speaker Beam[7] が提案されている．この手法では目的話者の補助情報と混合音声から得られた特徴量を統合し，混合音声から目的話者音声を抽出するマスクを推定するネットワークの学習を行う．抽出する音声が目的話者の元の音声に近くなるようにモデルが学習されるため，話者分離に有効な特徴表現が学習されることが期待できる．

そこで本研究では，妨害発話環境下のマルチモーダル音声認識モデルにおいて，音源分離モデルを併用した学習方式について検討する．音声認識モデルの学習において，文献 [7] のような音源分離モデルの学習方式を導入することにより，音源分離性能を向上させつつ音声認識に最適なパラメータが学習されることを期待する．本研究では学習方式として，音源分離モデルを用いて事前学習を行う方法と音声認識モデルとのマルチタスク学習を行う方法を検討する．

2 音声認識モデル

従来手法 [6] の概要を Fig. 1(a) に示す．従来手法では，CTC と Attention モデルを組み合わせて学習を行う Hybrid CTC/attention モデルを用いている．妨害発話環境下において，目的音声を分離する能力を強化するため，モデルの学習時に妨害発話に基づく補助損失関数を用いている．音声と画像を統合した特徴量を，目的話者の発話認識を行うための Target Encoder と，妨害話者の発話認識を行うための Interference Encoder に入力する．2 つの Encoder で処理された特徴量はそれぞれの Decoder に入力され，発話内容の認識結果を出力する．この時，目的話者の発話認識を行うための損失関数 L_{target} と妨害話者の発話認識を行うための損失関数 $L_{interference}$ は以

* Pre-training method for audio-visual speech recognition in interference speaker environment. by Ryota Tsunoda (Kobe University), Ryo Aihara (Mitsubishi Electric Corporation), Ryoichi Takashima, Tetsuya Takiguchi (Kobe University), Imai Yoshie (Mitsubishi Electric Corporation)

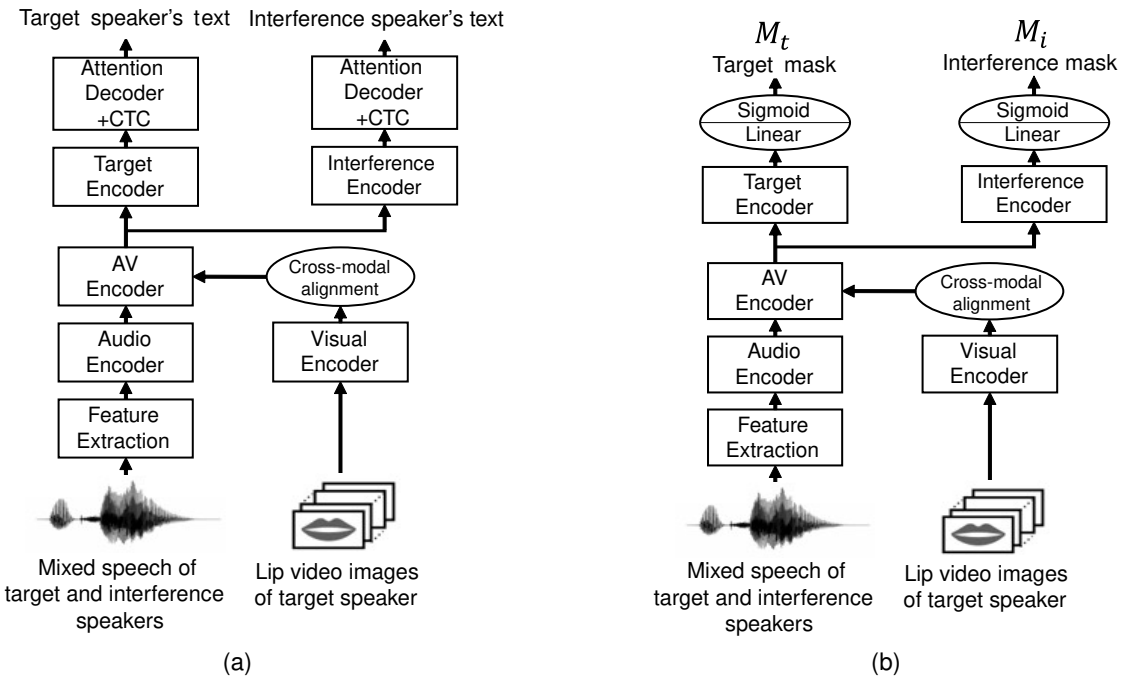

Fig. 1　(a) Speech recognition model, and (b) Speech separation model

下のように表される．

$$L_{traget} = \alpha L_{t\_ctc} + (1-\alpha) L_{t\_att} \quad (1)$$

$$L_{interference} = \alpha L_{i\_ctc} + (1-\alpha) L_{i\_att} \quad (2)$$

L_{x_ctc}, L_{x_att} はそれぞれ CTC 損失関数，Attention モデルにおける損失関数であり，x が t の時は目的話者，x が i の時は妨害話者を示す．また，α はマルチタスク学習における重みパラメータである．L_{target} に加え，$L_{interference}$ を導入することで妨害発話の認識精度を向上させることにより，AV Encoder で話者分離のための特徴表現を獲得し，目的音声と妨害音声を分離するようにモデルを学習する．この時，学習時に用いる全体の損失関数は以下の式で表される．

$$L_{rcg} = L_{traget} + L_{interference} \quad (3)$$

本研究ではこのモデルを Baseline とする．

3　提案手法

本研究では，音源分離モデルを併用して音声認識モデルの学習を行う手法を検討する．音源分離モデルの概要図を Fig. 1(b) に示す．Fig. 1(a) に示した音声認識モデルでは Target Encoder と Interference Encoder に続けて Hybrid CTC/attention decoder を用いて発話内容の認識結果を出力するが，音源分離モデルではその代わりに全結合層と Sigmoid 関数を用いて混合音声の特徴量から元の音声の特徴量を抽出するマスクの推定を行う．この時，目的話者と妨害話者のマスクを推定するための損失関数 L_{t_mse}, L_{i_mse} は以下のように表される．

$$L_{t\_mse} = MSE(X_t, M_t \odot Y) \quad (4)$$

$$L_{i\_mse} = MSE(X_i, M_i \odot Y) \quad (5)$$

ここで，X_t, X_i 及び Y はそれぞれ目的話者音声，妨害話者音声，混合音声の音響特徴量，M_t, M_i はそれぞれ目的話者，妨害話者の音声を抽出するためのマスク，$\odot$ は要素積，MSE は平均二乗誤差を示す．目的話者と妨害話者の元の音響特徴量を抽出することで，分離性能が向上するようにモデルが学習されることを期待する．また，このモデルの全体の損失関数は以下のように表される．

$$L_{spr} = L_{t\_mse} + L_{i\_mse} \quad (6)$$

本研究では，音源分離モデルの併用方法として 2 種類の学習方式を検討する．1 つ目は音源分離モデルを用いて事前学習を行う方法である．この方法では，音源分離モデルの学習を行った後に，全ての Encoder のパラメータを初期値として音声認識モデルの学習を行う．2 つ目は音源分離モデルと音声認識モデルのマルチタスク学習を行う方法である．この方法では，各 Encoder のパラメータを共有して，音源分離モデルと音声認識モデルを同時に学習する．マルチタスク学習

時の損失関数は以下のように表される．

$$L = \beta L_{spr} + (1-\beta) L_{rcg} \tag{7}$$

ここで，β は音源分離モデルと音声認識モデルのマルチタスク学習における重みパラメータである．

4 評価実験

4.1 実験条件

本研究では，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 6,913 文を発話している音声とビデオ映像で構成されている．TCD-TIMIT には単一話者による発話が収録されているため，本実験では二話者の音声を重畳することで，妨害発話環境音声を作成した．まず訓練データとして，TCD-TIMIT の 3,752 発話に対して，1 発話につき話者の異なる 7 発話をランダムに選んで重畳することで，26,264 発話分のデータを作成した．評価データは，TCD-TIMIT の 1,736 発話に対して，1 発話につき話者の異なる 1 発話をランダムに選んで重畳して作成した．この時，人間の聴覚特性に合わせた音の大きさの指標である Loudness[9] に従い，信号対雑音比を設定した．訓練データは 1 文毎に −10 dB から 10 dB の間でランダムに選択し，評価データは −10 dB から 10 dB の間にて 5 dB 刻みでそれぞれ固定したデータを作成した．

マルチモーダル音声認識モデルは ESPnet[10] を用いて，Hybrid CTC/Attention モデル [3, 11] の学習を行った．音声の入力特徴量には，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 次元の特徴量を使用した．画像の特徴量には，ビデオ映像から OpenFace[12] を用いて顔画像を検出後，唇領域を切り取って 36×36 にリサイズした 3 チャネルカラー画像を使用した．なお画像の特徴量には目的話者に対応したもののみが入力される．出力は英文字 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を加えた 31 次元とした．

以下にモデル構造の詳細について示す．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．統合処理を行う 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．また，Target Encoder，Interference Encoder には 320 次元の隠れ層を持つ 3 層の双方向 GRU を使用した．デコーダは 320 次元の隠れ層を持つ 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．Hybrid CTC/Attention におけるマルチタスク学習には CTC 損失関数の重みを 0.2，認識時の CTC の出力確率の重みを 0.2 に設定した．

Baseline 及び音源分離モデルと音声認識モデルのマルチタスク学習を学習する際，二話者混合音声を用いてモデルを学習する前に，まず TCD-TIMIT のオリジナルデータを用いて単一話者音声のモデルを学習し，それを初期モデルとして二話者混合音声モデルを学習した．この時，Interference Encoder は Target Encoder の重みを初期値とした．また, 音源分離モデルと音声認識モデルにおけるマルチタスク学習の重み β は 0.9 に設定した．

4.2 目的話者音声認識の性能比較

Table 1 に，各実験における文字誤り率 (Charactor Error Rates; CERs) を示す．Baseline はマルチモーダル音声認識モデルのみを用いて学習を行った場合である．本研究の提案手法である Separation pre-train，Separation MTL はそれぞれ音源分離モデルで事前学習を行う方法とマルチタスク学習を行う方法を示す．Baseline と Separation pre-train を比較すると，低雑音環境下では大きな差は見られなかったが，高雑音環境下では認識精度が向上し −10dB の時に 10.2%の相対的改善を示した．また，Baseline と Separation MTL を比較すると，Separation pre-train と同様に高雑音環境下における改善が見られ，−10dB の時に 9.1%の相対的改善を示した．この結果から，学習方式における違いに大きな差はなかったが，音源分離モデルの学習が音声の分離が困難である高雑音環境下における音声の分離性能向上に寄与し，結果的に認識精度が向上したと考えれられる．

4.3 妨害話者音声認識の性能比較

Baseline 及び提案手法において，Interference Encoder の出力を用いて妨害話者の発話内容を推定し，音声認識性能を評価した．Table 2 に評

Table 1　CERs of target speaker ASR

SNR of traget spk	Baseline	Separation pre-train	Separation MTL
10 dB	14.6	14.5	14.4
5 dB	16.8	16.3	16.2
0 dB	19.1	17.9	18.1
-5 dB	22.6	20.7	20.8
-10 dB	26.5	23.8	24.1

Table 2　CERs of the interference speaker ASR

SNR of interference spk	Baseline	Separation pre-train	Separation MTL
10 dB	23.1	23.4	22.5
5 dB	26.0	26.2	25.1
0 dB	31.5	29.9	28.8
-5 dB	39.3	36.1	35.5
-10 dB	49.4	45.3	44.7

価結果を示す．目的話者の音声認識精度と比較すると，妨害話者の補助情報は与えられていないため，認識精度は全体的に大きく低下していることがわかる．Baseline と比較すると Separation pre-train 及び Separation MTL では目的話者音声認識の結果と同様に特に高雑音環境下にて大きな改善を示した．この結果からも目的話者と妨害話者の音声を分離するための特徴表現が学習できていると考えられる．

5　おわりに

本研究では，マルチモーダル音声認識において，音源分離モデルを併用して学習を行うことで目的話者の音声認識精度の向上を図った．Baseline と比較して，音源分離モデルを併用して学習を行った場合，特に高雑音環境下における認識精度を改善する結果を示した．今後は，目的話者の発話内容をさらに高い精度で認識するために，単一話者音声での事前学習と音源分離モデルでの事前学習を効果的に組み合わせる手法について検討する．

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. TSP*, pp. 484–487, 2013.

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

[4] G. Sterpu *et al*, “Attention-based audio-visual fusion for robust automatic speech recognition,” in *Proc. ACM ICMI*, 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] 角田 他, “妨害発話に基づく補助損失を用いたマルチモーダル目的話者音声認識,” 日本音響学会秋季研究会講演論文集, pp. 1041–1044, 2021.

[7] T. Ochiai *et al*, “Multimodal speakerbeam: Single channel target speech extraction with audio-visual speaker clues,” in *Proc. Interspeech*, pp. 2718–2722, 2019.

[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] ITU-R BS.1770-4, “Algorithms to measure audio programme loudness and true-peak audio level,” 2012.

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

[11] 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.

[12] T. Baltrušaitis *et al*, “Openface 2.0: Facial behavior analysis toolkit,” in *Proc. FG*, pp. 59–66, 2018.