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*

☆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

人間は発話内容を理解する際に，様々な情報を統合的に利用している．特に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，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に対する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．従来のマルチモーダル音声認識モデルである AV Align [3, 4] では，背景雑音環境下において認識精度の改善を示した．しかし，複数の話者が同時に発話しているような妨害発話環境下にて目的話者の音声を認識しようとすると認識精度が大きく劣化するという課題がある．

シングルモーダルの音声認識において，目的話者の音声認識精度を向上させるための手法として，妨害発話に関する補助損失関数を用いた手法が提案されている [5]．この先行研究では，lattice-free maximum mutual information (LF-MMI) に基づく目的話者音声認識の損失関数に加えて，妨害発話に基づく補助損失関数を用いる手法を提案している．この時，目的話者の音声サンプルを i-vector に変換し補助情報として使用している．補助損失関数の導入により，目的音声と妨害音声を分離するようにモデルが学習され，その結果として認識精度が向上することを示している．

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，モデルの学習時に妨害発話に基づく補助損失関数と口唇動画像から発話内容の認識を行うための補助損失関数を併用することを検討する．

以下，第 2 章で本研究で Baseline として使用するモデルである AV Align と補助損失関数について紹介する．第 3 章で評価実験を行い，第 4 章で本稿をまとめる．

2 手法

2.1 AV Align

マルチモーダル音声認識の手法として，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 [3, 4] が提案されている．一般的に音声と動画像のフレームレートは異なるため，統合を行うためには音声・画像間の対応 (アライメント) を得る必要がある．AV Align では Cross-modal attention 機構を用いることで，音声・画像間のアライメントを自動的に学習することができる．

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 Encoder，Visual Encoder へ入力され，それぞれ高次の特徴量 $\boldsymbol{o}_A = \{o_{A_1}, o_{A_2}, \ldots, o_{A_N}\}$，$\boldsymbol{o}_V = \{o_{V_1}, o_{V_2}, \ldots, o_{V_M}\}$ に変換される．その後，Cross-modal attention 機構により AV Encoder で $\boldsymbol{o}_A$，$\boldsymbol{o}_V$ が統合され，$\boldsymbol{o}_{AV}$ が出力される．このモデルの概要を Fig. 1 に示す．本研究では，AV Align を Baseline とし提案手法の評価実験を行う．

2.2 補助損失関数

提案手法の概要図を Fig. 2 に示す．本研究では，CTC と Attention モデルを組み合わせて学習を行う Hybrid CTC/attention モデル [6, 7] を用いる．2.1 節の Cross-modal attention 機構にて音声と画像を統合した特徴量 $\boldsymbol{o}_{AV}$ を，目的話者の発話認識を行うための Target Encoder と，妨害話

*Interference speaker loss for audio-visual target speaker speech recognition. by Ryota Tsunoda (Kobe University), Ryo Aihara (Mitsubishi Electric Corporation), Ryoichi Takashima, Tetsuya Takiguchi (Kobe University), Imai Yoshie (Mitsubishi Electric Corporation)

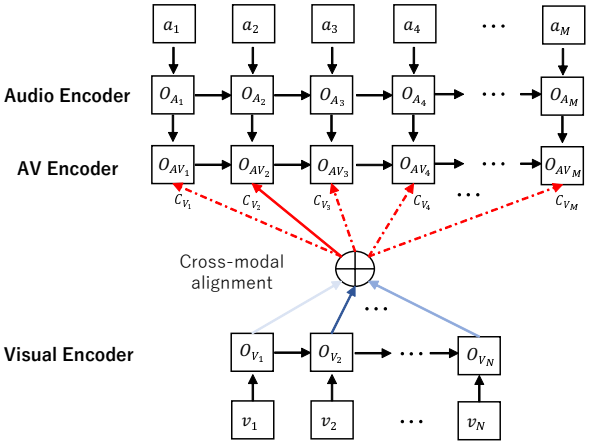

Fig. 1　Cross-modal attention mechanism

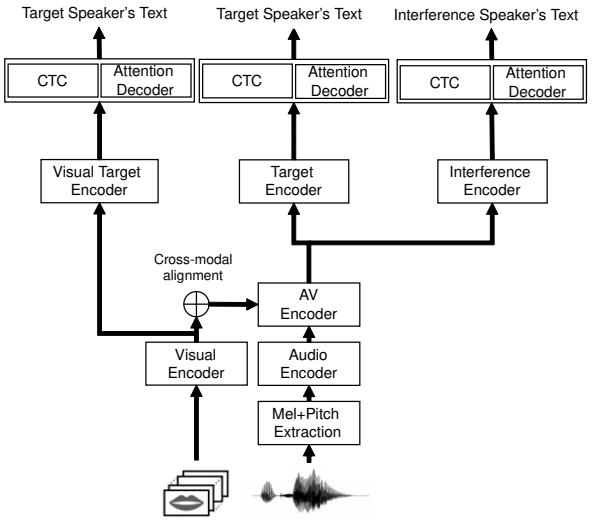

Fig. 2　Model architecture

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は以下のように表される．

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

ここで，L_{t_ctc}，L_{t_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，α はマルチタスク学習における重みパラメータである．

提案手法では L_{target} に加え，学習時に 2 つの補助損失関数を導入する．1 つ目は妨害話者の発話認識を行うための損失関数 $L_{interference}$ である．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が分離するようにモデルが学習されることを期待する．

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

2 つ目は口唇動画像から発話内容の認識を行うための損失関数 L_{visual} である．Visual Encoder 及び Visual Target Encoder で処理された特徴量が後段の Decoder に入力され発話内容の認識結果を出力する．口唇動画像から発話内容の認識精度を向上させることにより，Visual Encoder の学習を安定させ，目的話者の認識精度向上に寄与することを期待する．

$$L_{visual} = \gamma L_{v\_ctc} + (1-\gamma) L_{v\_att} \qquad (3)$$

L_{target}，L_{visual} は目的話者の音声認識精度を最大化するように働き，$L_{interference}$ は妨害話者の音声認識精度を最大化するように働く．本研究で学習時に用いる全体の損失関数は以下の式で表される．

$$L_{all} = \lambda_1 L_{traget} + \lambda_2 L_{interference} + \lambda_3 L_{visual} \qquad (4)$$

λ_1，λ_2，λ_3 はそれぞれの損失関数における重みパラメータである．

3　評価実験

3.1　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 は単一話者による発話が収録されており，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 End-to-End 音声認識ツールキットである ESPnet [10] を用いて，Hybrid CTC/Attention モデル [6, 7] の学習を行った．音声の入力特徴量には，23 次元のメ

Table 1　Character Error Rates (CER) of each experimental condition

Model	CER[%]				
	10 dB	5 dB	0 dB	-5 dB	-10 dB
Audio only	29.7	38.9	49.6	57.2	60.5
Baseline	18.4	21.1	23.9	27.5	30.6
+visual loss	21.6	24.3	27.2	30.5	33.3
+interference loss	15.7	17.8	20.3	23.1	26.4
+visual loss	17.4	19.4	22.0	24.5	27.9

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 [11] 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 [4] 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 及び Visual Target 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 0.5，認識時の CTC の出力確率の重みを 0.5 に設定した．また，(4) 式における損失関数の重みパラメータ λ_1，λ_2 及び λ_3 は全て 1.0 とした．

モデルを学習する際，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 のオリジナルの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．

3.2　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 Rate; CER) を示す．Baseline は Fig. 2 に示すマルチモーダル音声認識モデルの学習において $L_{interference}$，L_{visual} を使用しない場合，すなわち L_{target} のみを用いて学習を行った場合であり，従来の AV Align に相当する．目的話者の補助情報がない音声のみ (Audio only) の場合，目的話者と妨害話者の音声を区別することができないため，認識が困難であると考えられる．音声のみの結果とマルチモーダル音声認識を行った場合を比較すると，目的話者の補助情報として口唇動画像を使用することで，認識精度を改善することができている．このことから，目的話者の音声認識において補助情報の有無は認識精度に大きく影響を与えると考えられる．また，Baseline と $L_{interference}$ を使用した場合の結果を比較すると，相対比で 13.7%から 16.0%の改善を達成した．このことから，妨害発話環境下において妨害話者の発話認識を行うための損失関数を用いることは目的話者の発話認識の精度を向上させるために有効であると言える．その一方で，Baseline と L_{visual} を使用した場合の結果を比較すると，認識精度は低下した．さらに，L_{visual} と $L_{interference}$ の両方を使用した場合 Baseline よりも認識精度が改善しているが，$L_{interference}$ のみを使用した場合より低下している．一般的に，音声から発話内容の認識を行った場合と比較して口唇動画像から発話内容の認識を行った場合認識精度は下がることから，L_{visual} を加えたことにより Visual Encoder における特徴表現の学習を妨げたことが原因と考えられる．

Table 2 Experimental results of the target speaker ASR and the interference speaker ASR

SNR of traget spk	SNR of interference spk	target spk	interference spk
10 dB	-10 dB	15.7	56.0
5 dB	-5 dB	17.8	47.1
0 dB	0 dB	20.3	38.7
-5 dB	5 dB	23.1	31.5
-10 dB	10 dB	26.4	27.0

3.3 妨害話者音声認識の性能評価

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低いことがわかる．しかしその一方で，発話内容を認識する話者の補助情報がない音声のみの場合 (Table 1 の Audio only を参照) と比較すると高い精度を示している．この結果から，目的話者の補助情報を入力することで目的話者と妨害話者の音声を分離するための特徴表現を学習でき，その結果として妨害話者の音声認識もある程度可能になっていると考えられる．

4 おわりに

本研究では，マルチモーダル目的話者音声認識において，学習時に複数の補助損失関数を使用する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，妨害話者の発話認識を行うための損失関数を用いた手法では全ての信号対雑音比において認識精度を改善する結果を示した．また，妨害話者の発話認識を行った場合，音声のみを使用した場合よりも良い結果を示した．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，より優れた特徴表現を獲得するための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. International Conference on Telecommunications and Signal Processing (TSP)*, pp. 484–487, 2013.

[3] 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.

[4] 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.

[5] N. Kanda *et al*, “Auxiliary interference speaker loss for target-speaker speech recognition,” in *Proc. Interspeech*, pp. 236–240, 2019.

[6] 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.

[7] 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.

[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,” 2015.

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

[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.