クロスモーダル知識蒸留に基づく Lip reading のための教師なしドメイン適応*

◎高島悠樹 (神戸大), 相原龍 (三菱電機), 高島遼一, 滝口哲也,
有木康雄 (神戸大), 村山修 (三菱電機)

1 はじめに

Lip reading は，発話者の唇や顔，舌の動きから発話内容を視覚的に理解する技術である．例えば，聴覚障害者は周囲の音を聞くことが難しいため，コミュニケーション手段の 1 つとして lip reading を用いる場合がある．人間は発話内容を理解する際，種々の情報を統合的に利用している．音声が聞き取り難い場合，発話者の顔，特に唇の動きに注目して発話内容を理解しようとし，逆に，唇の動きと音声が不一致の場合，唇の動きに影響されて発話内容を誤って理解してしまうこともある．これは，McGurk effect (マガーク効果) と呼ばれ，音韻知覚が音声の聴覚情報のみで決まるのではなく，唇の動きといった視覚情報からも影響を受けることが報告されている [1]．そのため，発話内容の理解において，音声と唇の動きの統合的利用は極めて重要である．

近年，深層学習に基づくモデルが盛んに研究され，大量の学習データを用意できる環境において，音声と唇動画像を併用して発話認識を行うマルチモーダル音声認識や lip reading の性能を向上させている [2, 3, 4, 5]．一般に，学習データの環境とユーザの利用環境のミスマッチにより，実用シーンでは認識性能が低下する．しかし，あらゆる環境のデータを収集することは容易ではなく，学習したモデルを利用環境に応じて適応させることが求められる．この技術はドメイン適応と呼ばれる．

ドメイン適応の目的は，元ドメインで学習されたモデルを，出来る限り少量の適応データを用いて新しい目標ドメインに適応することである．特に，適応データのラベル情報を全く使用できない場合は，教師なしドメイン適応 (unsupervised domain adaptation; UDA) と呼ばれる．様々な UDA の手法 [6, 7, 8] が提案されているが，これらの多くは適応データと学習データが同じクラス集合に属していることを前提としている．そのため，未知のクラスに属するデータを適応データとして使用することができない．より実用的なドメイン適応のために，このような未知クラスのデータを利用できる適応手法が求められる．そこで本研究では，未知クラスのデータを用いた UDA の手法について検討する．同様の状況を想定した手法 [8, 9] が提案されているが，従来の適応手法が使えないため多くの課題が残っているタスクとなっている．

本研究では，lip reading をタスクとして，クロスモーダルな知識蒸留 (knowledge distillation; KD) に基づくアプローチを提案する．KD [10] はモデル圧縮手法として提案されたものであり，学習済みのモデル (教師モデル) の振る舞いを模倣するように対象モデル (生徒モデル) を学習する．KD は教師モデルの知識を生徒モデルへ転移させていると解釈でき，この性質を利用してドメイン適応など，他のタスクへも応用されている [11, 12, 13]．本稿では，生徒モデルを lip reading モデル，教師モデルを音声認識モデルとして，クロスモーダル KD を行う．ここで，lip reading モデルの学習および適応のために，唇画像データだけでなく音声データも用いる．まず音声認識モデルを学習し，その中間出力を用いて lip reading モデルを学習する．一般に発話認識において，音声は唇の動きよりも，より多くの情報を持っているとされ，音声認識の精度は lip reading よりも高くなる．そのため，音声認識モデルの出力は強力な教師信号となると期待される．

本研究では，クロスモーダル KD に基づくドメイン適応を提案し，さらに未知クラスのデータを用いた適応も可能であることを示す．通常の KD は教師モデルと生徒モデルの最終層の出力確率分布間の距離を最小化するが，未知クラスデータの場合，モデルの出力層には対応するクラスラベルが存在しない．そのため，KD を行うことができない．この問題を解決するために，最終層ではなく中間層を用いた KD を行う．モデルの中間層にサブクラス表現が現われていると仮定し，このサブクラス表現を用いてドメイン適応を行う．クラスを単語とした場合，サブクラスは sub-word や音素のようなものに相当し，これを用いることで未知クラスに属するデータであってもドメイン適応への使用が可能となる．単語認識実験により，提案手法が UDA 性能を向上させ，既知クラスを用いたドメイン適応と同等の性能を持つことを示す．

以下，第 2 章で提案手法を述べる．第 3 章で評価

*Unsupervised Domain Adaptation for Lip reading Based on Cross-modal Knowledge Distillation, by Yuki Takashima (Kobe University), Ryo Aihara (Mitsubishi Electric Corporation), Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki (Kobe University), Shu Murayama (Mitsubishi Electric Corporation)

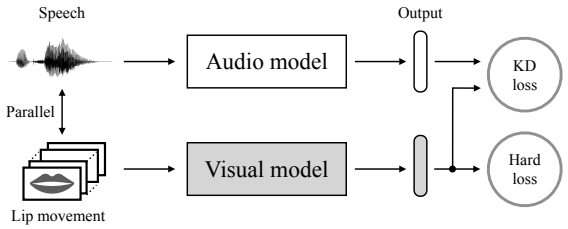

Fig. 1　Basic procedure of cross-modal knowledge distillation.

実験を行い，第 4 章で本稿をまとめる．

2　提案手法

本稿では，lip reading のための，未知クラスのデータを用いた UDA を目的とする．単語認識を想定しており，顔動画像をモデルへ入力し，対応する単語クラスを推定する．提案手法において，マルチモーダルデータはモデルの学習時と適応時に用い，評価時は顔動画像のみを入力とする．

2.1　クロスモーダル KD

Fig. 1 にクロスモーダル KD の概要を示す．ここで，音声と唇動画像はそれぞれ同一発話から抽出される．また，出力は単語ラベルである．まず，教師モデルである音声モデルを，入力音声とその正解単語ラベルを用いて，クロスエントロピー損失 (Hard loss) により学習する．次に，音声モデルを用いて画像モデルを学習する．音響特徴量 x_{aud} と画像特徴量 x_{vis} が与えられた時，KD ロスは以下の式で定義される．

$$-\sum_{l} p_{\mathrm{aud}}(l|x_{\mathrm{aud}}) \ln p_{\mathrm{vis}}(l|x_{\mathrm{vis}}). \tag{1}$$

ここで，$p_{\mathrm{vis}}(l|x_{\mathrm{vis}})$ と $p_{\mathrm{aud}}(l|x_{\mathrm{aud}})$ はそれぞれ，ラベル l に対する，画像モデルと音声モデルの出力確率を表す．この時，音声モデルのパラメータは固定する．この損失関数により，画像モデルは音声モデルの出力を模倣するように学習される．実際には，安定的な学習のために，正解ラベルを用いたクロスエントロピー損失との重み付け和を用いる．Li ら [13] は，音声認識モデルとマルチモーダル音声認識モデルの間の KD により，認識性能が向上することを示している．そのため，音声認識モデルと lip reading モデル間の KD も性能向上に貢献すると期待される．

2.2　クロスモーダル KD に基づく UDA

本研究では，適応データの属するクラスが元ドメインのどのクラスにも属していないとする．

提案手法では Fig. 2 に示すように，エンコーダと識別器からなるモデルを用いる．音声モデルと画像モデルに対するエンコーダは以下のように定義される．

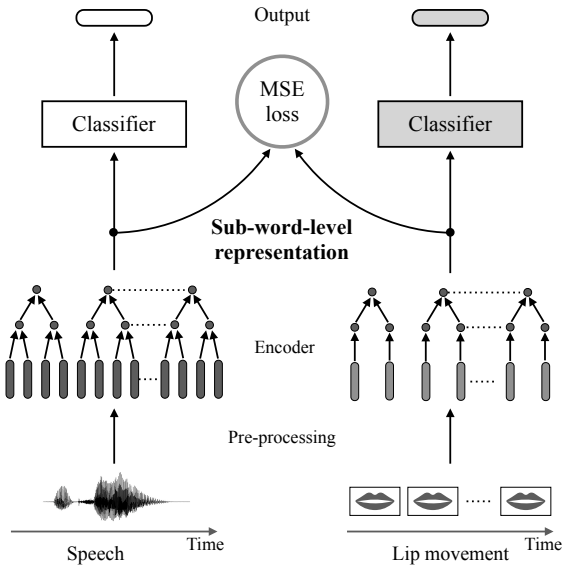

Fig. 2　Overview of our proposed UDA.

$$\boldsymbol{h}^a = f_{\mathrm{aud}}(\boldsymbol{a}), \tag{2}$$

$$\boldsymbol{h}^v = f_{\mathrm{vis}}(\boldsymbol{v}), \tag{3}$$

ここで，$\boldsymbol{a} = (a_1, ..., a_t, ..., a_{T_a})$ と $\boldsymbol{v} = (v_1, ..., v_{t'}, ..., v_{T_v})$ はそれぞれ，音響特徴量系列と画像特徴量系列を表す．$\boldsymbol{h}^a = (h_1^a, ..., h_u^a, ..., h_U^a)$ と $\boldsymbol{h}^v = (h_1^v, ..., h_u^v, ..., h_U^v)$ はそれぞれ，音声と画像の潜在特徴系列である．また，a_t, $v_{t'}$, $h_u^a \in \mathbb{R}^d$, $h_u^v \in \mathbb{R}^d$ はそれぞれ，t フレーム目の入力音響特徴量フレーム，t フレーム目の入力画像特徴量フレーム，および，それぞれのモダリティに対する d 次元のエンコードされた潜在特徴量を表す．T_a, T_v, $U \leq \min(T_a, T_v)$ はそれぞれ，音響特徴量と画像特徴量のフレーム数，および，潜在特徴のフレーム数である．潜在特徴系列のタイムステップ数 U はモダテリィ間で共通とする．識別器は全結合層であり，最終層の次元数は単語数である．

提案する UDA では，元ドメインで学習した画像モデルをクロスモーダル KD により目標ドメインに適応する．適応ステップでは，画像モデルのパラメータは以下に示す潜在表現間の mean squere error (MSE) を最小化するように更新される．

$$\mathcal{L}_{MSE}(\mathcal{D}) = \mathbb{E}_{\{\boldsymbol{v},\boldsymbol{a},\boldsymbol{y}\}\sim\mathcal{D}}[||\boldsymbol{h}^a - \boldsymbol{h}^v||_2^2]. \tag{4}$$

ここで，$\mathcal{D}$ は音響特徴量および画像特徴量，単語ラベルの三つ組のデータ集合を表す．また，ラベル $\boldsymbol{y}$ は使用しない．通常の KD と異なり，中間層の潜在表現を用いて蒸留を行う．識別器は単語ラベルを推定するためフレームレベルの情報は消失しており，単語レベルの特徴表現が伝搬していると考えられる．そのた

め，最終出力層および識別器中間層を用いる KD では未知クラスのデータを扱うことができない．一方で，エンコーダはフレームレベルの情報を残しており，それは sub-word や音素のような表現に対応すると考えられる．これらは単語非依存であるため，提案手法は未知クラスのデータを用いた UDA を行うことが可能であると考えられる．

2.3 学習方法

元ドメインと目標ドメインのデータ集合をそれぞれ，$\mathcal{D}_{\mathrm{src}}$，$\mathcal{D}_{\mathrm{trg}}$ とする．学習ステップでは，クロスモーダル KD に基づいて元ドメイン画像モデルの学習を行う．まず初めに，正解ラベルに対するクロスエントロピー損失を用いて，以下の損失関数により音声モデルを学習する．

$$\mathbb{E}_{\{\boldsymbol{v},\boldsymbol{a},\boldsymbol{y}\}\sim\mathcal{D}_{\mathrm{src}}}[-\log(g_{\mathrm{aud}}(\boldsymbol{y}|\boldsymbol{h}^a))]. \tag{5}$$

ここで，$g_{\mathrm{aud}}(\boldsymbol{y}|\boldsymbol{h}^a)$ は，潜在特徴 $\boldsymbol{h}$ から推定された音声モデルの識別器によるラベル $\boldsymbol{y}$ の出力確率を表す．また，画像特徴量 $\boldsymbol{v}$ は使用しない．画像モデルは，音声モデルを用いたクロスモーダル KD 損失と正解ラベルに対するクロスエントロピー損失により，以下の損失関数により学習される．

$$\begin{aligned}&\mathcal{L}_{MSE}(\mathcal{D}_{\mathrm{src}})+\mathcal{L}_{CE}(\mathcal{D}_{\mathrm{src}})\\&=\mathcal{L}_{MSE}(\mathcal{D}_{\mathrm{src}})+\mathbb{E}_{\{\boldsymbol{v},\boldsymbol{a},\boldsymbol{y}\}\sim\mathcal{D}_{\mathrm{src}}}[-\log(g_{\mathrm{vis}}(\boldsymbol{y}|\boldsymbol{h}^v))].\end{aligned} \tag{6}$$

ここで，$g_{\mathrm{vis}}(\boldsymbol{y}|\boldsymbol{h}^v)$ は画像モデルの識別器による出力確率である．

適応ステップでは，未知クラスのデータを用いて画像モデルの適応を行う．学習の安定化のために，式 (4) だけでなく，正解ラベルを持つ元ドメインデータに対する損失も計算する．この項は，音声モデルの潜在特徴分布に対する過適合を防ぐ役割を持つ．提案する未知クラスのデータによる UDA は以下の損失を最小化する．

$$\begin{aligned}\mathcal{L}=\mathcal{L}_{MSE}(\mathcal{D}_{\mathrm{trg}})+\alpha\mathcal{L}_{MSE}(\mathcal{D}_{\mathrm{src}})\\+(1-\alpha)\mathcal{L}_{CE}(\mathcal{D}_{\mathrm{src}}).\end{aligned} \tag{7}$$

ここで，α は重み係数であり，本項では 0.5 を使用する．

3 評価実験

3.1 実験条件

提案手法は，単語認識タスクにより評価した．Audio-visual 音声認識データセット lip reading in the wild (LRW) [4] を使用した．LRW は，イギリスのテレビ放送から抽出した音声と唇動画像からなるデータセットであり，数百人の話者が発話した 500 単語 (各単語 1,000 発話程度) から構成される．各発話は 29 フレーム (1.16 秒) に切り出されている．本稿では全 500 単語を，既知クラス 400 単語と未知クラス 100 単語に分割した．既知クラスのデータは元ドメインモデルの学習，および評価に使用される．提案手法は，既知クラスデータに対する適応と未知クラスデータに対する適応の両方で評価を行った．元ドメインとして，データベースのデータをそのまま使用し，目標ドメインのデータとして，顔画像の輝度値を変化させた (音声には一切の加工を行わない) ものを使用した．学習データとして元ドメインの既知クラス単語各 500 発話 (合計 200,000 発話)，適応データとして目標ドメインの未知クラス単語各 250 発話 (合計 25,000 発話) を使用した．評価は，既知クラス 400 単語の識別タスクにより行った．評価データとして，LRW の評価セットに含まれる既知クラス分の発話 (合計 20,000 発話) を使用し，画像に対して適応データと同様の加工を行った．比較のために，既知クラス適応データとして，学習データとは異なる各単語 250 発話を選択した．

音声モデルの入力特徴量として，1 発話分の 40 次元対数メルフィルタバンク係数と Δ，ΔΔ をチャネル方向へ重ねたものを用いた．STFT におけるフレーム幅，シフト幅はそれぞれ 25ms，5ms としたため，1 発話 116 フレームとなる．画像モデルの入力特徴量として，各フレームをグレースケールへ変換し，28 フレームをチャネル方向へ重ねたものを用いた．音声エンコーダは，隣接する 2 フレームを入力とする畳み込み層を 3 層積層し，タイムステップごとに 128 次元の全結合層を配置した．畳み込み操作では，フレーム方向のストライドを 2 とし，層を経るごとにタイムステップ数が半分となる構造を採用した．画像エンコーダは，フレームごとの畳み込み層の後に，隣接する 2 フレームを入力とする畳み込み層 (ストライド 2) を 1 層積層し，タイムステップごとに 128 次元の全結合層を配置した．識別器は 3 層の全結合 (4,096 → 4,096 → 400) を用いた．モデルの最適化には Adam を用いた．バッチサイズは 24，学習率は 1e-4 とした．学習時は開発データを用いた早期終了 (early stopping) を行い，適応時のエポック数は 10 とした．

3.2 結果と考察

まず，クロスモーダル KD の評価を行う．ここでは，元ドメインのデータのみを扱い，ドメイン適応は考慮しない．実験結果を Table 1 に示す．“Baseline” は，顔画像のみを用いて学習されたベースラインモデルを示す．提案モデルでは KD のために中間層の

Table 1 Word recognition accuracy [%] for each method on the source domain. # Utterance/word shows the number of utterances per word used to train the model.

	# Utterance/word	
Model	250	500
Baseline	48.21	54.62
Proposed	50.06	55.07

Table 2 Word recognition accuracy [%] for each method. First column shows a domain of the evaluation data.

Eval. domain		Baseline	Prop.
Source		54.62	55.07
Target	No adap.	39.19	42.84
	Known class adap.	44.10	52.22
	Unknown class adap.	—	52.04

次元数を削減しているため，顔画像のみを用いて学習すると大きな性能劣化を引き起こす．そのため，次元削減を行わないモデルをベースラインモデルとした．表より，各単語 250 発話を使用した場合，提案モデルはベースラインモデルと比べて 3.57%の相対性能向上を達成した．この結果は，音声認識モデルの出力が強力な教師として働くことを示している．また，KD は正則化の効果もあるため，学習データが少ない場合 (各単語 250 発話) により効果が得られた．

次に，提案する未知クラスのデータに対する UDA の評価を行った．ベースライン適応手法として，適応データを元ドメインモデルへ入力し，推定されたラベル情報を用いてモデルの再学習を行った．Table 2 に各手法の単語認識率を示す．まず，既知クラス適応において，提案手法は従来手法よりも優れた認識率を示した．適応なしと比べて 12.53%，従来の適応手法と比べて 21.88%の相対性能向上を達成した．さらに，未知クラス適応においても，提案手法は 21.48% の相対性能向上を達成した．これらの結果から，提案する中間層 KD は，識別クラスに依存しないサブクラス表現を転移可能であることを示している．さらに，そのような表現を用いることで，未知クラスのデータであっても，効果的なドメイン適応を可能にする．

4 おわりに

本研究では，未知クラスデータによるドメイン適応のための中間層 KD アプローチを提案し，音声認識モデルの知識を効果的に lip reading モデルへ転移させることができることを示した．評価実験により，適応データが既知クラスであっても未知クラスであっても，同等の適応性能を達成した．本稿では人工的に目標ドメインのデータを生成したが，今後はより実環境に近いドメインに対する適応の評価を行う．

参考文献

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," Nature, vol. 264, pp. 746–748, 12 1976.

[2] A. Verma *et al.*, "Late integration in audio-visual continuous speech recognition," in *Proc. IEEE ASRU*, vol. 1, pp. 71–74, 1999.

[3] K. Palecek 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.

[4] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. ACCV*, pp. 87–103, 2016.

[5] J. S. Chung *et al.*, "Lip reading sentences in the wild," in *Proc. IEEE CVPR*, pp. 3444–3453, 2017.

[6] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. ICML*, pp. 1180–1189, 2015.

[7] M. Ghifary *et al.*, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proc. ECCV*, vol. 9908, pp. 597–613, 2016.

[8] K. Saito *et al.*, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE CVPR*, pp. 3723–3732, 2018.

[9] P. P. Busto and J. Gall, "Open set domain adaptation," in *Proc. IEEE CVPR*, pp. 754–763, 2017.

[10] G. Hinton *et al.*, "Distilling the knowledge in a neural network," in *Proc. NIPS Deep Learning Workshop*, 2014.

[11] T. Asami *et al.*, "Domain adaptation of dnn acoustic models using knowledge distillation," in *Proc. IEEE ICASSP*, pp. 5185–5189, 2017.

[12] G. Chen *et al.*, "Learning efficient object detection models with knowledge distillation," in *NIPS*, pp. 742–751, 2017.

[13] W. Li *et al.*, "Improving audio-visual speech recognition performance with cross-modal student-teacher training," in *Proc. IEEE ICASSP*, pp. 6560–6564, 2019.