Lip reading のためのクロスモーダル Teacher-Student 学習 *

◎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

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]．このように人間による発話内容の理解には，唇の画像と音声の情報の統合的利用が極めて重要である．

音声に唇動画像を併用して発話認識を行う手法として，マルチモーダル (audio-visual) 音声認識が研究されている [2, 3]．マルチモーダル音声認識は，車載カメラ映像など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．また，監視カメラに収録された会話映像のように音声が聞き取りにくい場合であっても，発話内容の分析が可能であり，犯罪の防止や抑止に繋がると考えられる．一般に，唇画像の持つ情報量は音声よりも少なく，マルチモーダル音声認識の性能は唇画像のみからの認識性能に強く影響される [4]．マルチモーダル音声認識の性能向上のために，唇の動きのみから発話内容を理解する，リップリーディング (読唇) の性能を向上させる必要があると考えられる．

リップリーディングの応用例として，難聴者の支援システムがある．難聴者は耳で音を聞くことができないため，正確な発音をすることが難しく，発話スタイルが健常者と異なる．そのため，一般的な音声認識システムでは彼らの音声を認識することは難しい．彼らの中には訓練により意図した発話の唇の形状を作ることが可能な方もおり，聴覚障害者にとってリップリーディングは重要なコミュニケーション手段の一つとなっている．しかし，一般にリップリーディングは難しいため，それを機械により行うことで難聴者と周囲の人とのコミュニケーションを支援することができる．

音声認識と比べてリップリーディングは難しく，認識性能は大きく劣化する．そこで本研究では，転移学習の 1 つである teacher-student (TS) 学習を用いて，リップリーディングの性能を向上させる手法を提案する．

TS 学習は knowledge distillation [5] とも呼ばれる，ニューラルネットワークの学習手法の 1 つである．一般に大規模なニューラルネットワークは優れた性能を示すが，モバイル端末など，計算資源が限られた環境では，大規模なネットワークを動作させることは難しい．そこで，学習済みの大規模なモデルを圧縮して小規模なモデルを得る手法として TS 学習が用いられる．TS 学習では，大規模なモデル (教師モデル) の出力を教師信号として小規模なモデル (生徒モデル) を学習する．生徒モデルは教師モデルの出力を模倣するように学習されるため，大規模・高性能な教師モデルの知識を転移させることができる．TS 学習は，モデル圧縮 [6] や雑音に頑健な音響モデルの構築 [7] に応用され，優れた性能を示している．本研究では，TS 学習を異なるモダリティ間での知識転移に応用する．

文献 [8] では，RGB 画像から深度画像への知識転移が提案されている．TS 学習を一般化し，任意の層での知識転移として定式化した．文献 [9] では，TS 学習を用いたマルチモーダル音声認識が提案されている．モデルの学習のために，audio-visual データセットだけでなく，大規模な音声データセットを使用することで，性能向上を実現した．本研究では，音声認識モデルを用いた TS 学習により，リップリーディングモデルを学習する．TS 学習により音声の知識を効果的に転移させられるため，リップリーディングの性能向上が期待できる．文献 [9] との違いとして，本研究では追加の音声データセットを使用せず，audio-visual データセット LRW (Lip Reading in the Wild [10]) に含まれている音声と唇動画像のみを使用する．実用シーンを想定すると大量の学習データを用意することは難しく，少量データによるモデル学習・適応が望まれる．そこで，学習データ量を変化させた場合の性能について評価を行う．

以下，第 2 章で公開されている audio-visual データセットを紹介する．第 3 章で TS 学習について簡潔に述べ，第 3 章で提案手法を述べる．第 4 章で評価実験を行い，第 5 章で本稿をまとめる．

---

* Cross-modal Teacher-Student Learning for Lip reading, by Yuki Takashima (Kobe University), Ryo Aihara (Mitsubishi Electric Corporation), Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki (Kobe University), Shu Murayama (Mitsubishi Electric Corporation)

Table 1　Statistics of audio-visual speech recognition databases.

Name	Type	# Vocabulary	# Speaker	Notes
GRID [11]	Words	51	34	Frontal face
CUAVE [12]	Digits	10	36	Frontal and side face
AVICAR [13]	Digits, Sentences	—	100	In-car
LRW [10]	Words	500	1,000+	TV
LRS [14]	Sentences	17,428	Thousands of speakers	TV

2　Audio-Visual データセット

Table 1 に，公開されている audio-visual データセットの一部を示す．GRID [11] と CUAVE [12] はコマンドベースであり，比較的単純なタスク設定である．AVICAR [13] は，車載カメラの映像を想定しており，4 つのカメラ・7 つのマイクロホンにより収録されている．音響的な雑音として窓の開閉と速度，画像への雑音として顔向きや日照の変化がある．LRW [10] は，イギリスのテレビ放送 BBC によって収録された映像から抽出した単語発話から構成されている．千人以上の話者による発話が収録されており，大規模なデータセットである．テレビ番組から抽出されているため，様々な顔向きや話者を含む．LRS [14] もテレビ放送から抽出されたデータセットである．LRS は LRW と異なり，文単位認識をタスクとして設計されている．

実用シーンでは，ある特定の環境に対して，必ずしも大量のデータを用意することができないことが考えられる．そこで，少量データによるモデルの学習・適応技術が求められる．本研究では，音声認識モデルが学習した知識を用いた TS 学習により，リップリーディングモデルの性能向上を試みる．

3　Teacher-Student 学習

Teacher-student (TS) 学習はニューラルネットワークのモデル圧縮のための学習手法の 1 つであり，大規模なモデルから小規模なモデルを学習するために用いられる．TS 学習では，教師モデルと生徒モデルが存在し，正解ラベルではなく，あらかじめ学習された教師モデルの出力を教師信号として生徒モデルを学習する．

TS 学習のフレームワークでは，まず，正解ラベルを用いて教師モデルを学習する．次に，教師モデルの出力を教師信号として，クロスエントロピー損失を用いて生徒モデルを以下の式を用いて学習する．

$$\mathcal{L}_{\mathrm{TS}} = -\sum_{l} p_{\mathrm{tea}}(l|x) \ln p_{\mathrm{stu}}(l|x). \tag{1}$$

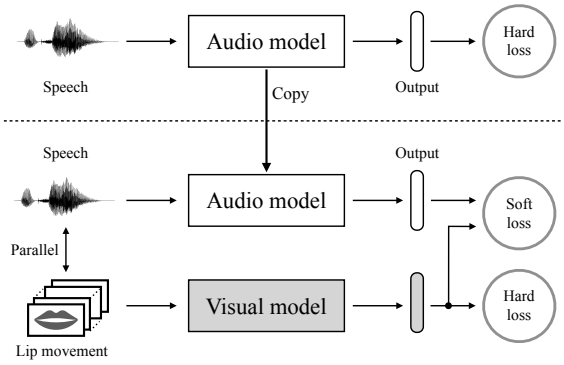

Fig. 1　Overview of our proposed method. “Hard loss” and “Soft loss” indicate the cross-entropy using the correct label and the TS loss using the outputs from the teacher model, respectively.

ここで，$p_{\mathrm{tea}}(l|x)$ と $p_{\mathrm{stu}}(l|x)$ はそれぞれ，入力 x に対するラベル l の教師モデルと生徒モデルの事後確率を表す．実際の学習では，正解ラベルに対する損失と式 (1) の損失を線形補間したものを目的関数として用いる．

4　提案手法

Fig. 1 に提案手法の概要を示す．提案手法では，まず，音声から単語認識を行うネットワークを学習する．入力音響特徴量 x_{aud}，単語ラベル l に対する音声ネットワークの出力を $p_{\mathrm{aud}}(l|x_{\mathrm{aud}})$ とした時，正解ラベル l に対するクロスエントロピー損失は以下の式で表される．

$$-\ln p_{\mathrm{aud}}(l|x_{\mathrm{aud}}). \tag{2}$$

そして，学習された音声認識ネットワークを教師モデルとして TS 学習によりリップリーディングネットワークの学習を行う．損失は以下の式で表される．

$$\begin{aligned} &-(1-\lambda)\ln p_{\mathrm{vis}}(\hat{l}|x_{\mathrm{vis}}) \\ &\quad -\lambda\sum_{l} p_{\mathrm{aud}}(l|x_{\mathrm{aud}}) \ln p_{\mathrm{vis}}(l|x_{\mathrm{vis}}). \end{aligned} \tag{3}$$

ここで，$p_{\mathrm{vis}}(l|x_{\mathrm{vis}})$ は入力唇画像 x_{vis} に対するラベル l の事後確率を表す．また，$\hat{l}$，λ はそれぞれ，入力に対する正解ラベル，線形補間重みを表す．ここで，TS 学習の損失計算時には，入力唇動画像に対応する音声を教師モデルへ入力する．一般に，音声と比べて，唇画像の持つ情報量は少ない．TS 学習により，リップリーディングの性能向上が期待できる．

正解ラベルを用いた損失では，正解単語であるということしか学習できないが，TS 学習による損失により，それぞれの単語らしさを学習することができる．つまり，ある単語と似ているが，ある単語とは似ていない，というような情報を学習できる．また，TS 学習は，厳密だが学習が難しい正解ラベルとの損失でなく，誤りも含むが学習が容易な教師モデルの出力を用いるため，正則化の効果があると考えられる．

5 評価実験

5.1 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 LRW [10] 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．文章中から単語を切り出しているため，発話の始終端には当該単語と異なる音を含む．本稿では，各単語 500 発話 (“Cond. 1”) および 250 発話 (“Cond. 2”) からなる 2 つのサブセットを学習データとして使用した．開発データおよび評価データは，各単語 50 発話からなる 25,000 発話から構成される．

音声認識ネットワークおよびリップリーディングネットワークの構造を Table 2，3 に示す．音声認識ネットワークの入力特徴量として，1 発話分の 40 次元ログ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 であるため，1 発話 116 フレームとなる．リップリーディングネットワーク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，1 発話分 29 フレームをチャネル方向へ重ねたものを用いた．学習率 0.003，モーメンタム 0.9，重み減衰 0.001 の確率的勾配法を用いてモデルを学習した．バッチサイズは 24 とした．開発データを用いた早期終了 (early stopping) を行った．

5.2 結果と考察

まず，TS 学習を用いない，音声あるいは画像のユニモーダル単語認識実験を行なった．実験結果を Table 4 に示す．音声のみによる認識 (音声認識) は，画像のみによる認識 (リップリーディング) に比べ，平均で 48.68% 高い認識性能を示した．これは，音声と唇画像が持つ情報量の乖離が大きいことを示す．また，学習データ量が半減した場合，音声のみによる認識では 2.50% の相対誤差だが，画像のみによる認識では 17.31% も低下した．このように，唇画像の持つ情報量は音声に比べて少ないため，正解単語を認識することは難しい．

Fig. 2 に TS 学習を用いたリップリーディングの実験結果を示す．音声モデルを用いた TS 学習により，認識性能が向上することが分かる．これは，音声認識ネットワークが学習した知識を，リップリーディングネットワークへ適切に蒸留できたためだと考えられる．また，Cond. 1 は TS 学習重み 0.5 の時に 1.14%，Cond. 2 は TS 学習重み 0.6 の時に相対的に最大で 4.94%の性能向上を達成した．学習データが少ない方

Table 2 Network architecture of the audio model.

Layer index	Operation
0	40× 116 × 3 Input
1	5×7 conv, 96, ReLU, 2×2 max-pool
2	5×7 conv, 256, ReLU, 2×2 max-pool
3	5×7 conv, 512, ReLU
4	4,096 dense, ReLU
5	4,096 dense, ReLU
6	500 dense, softmax

Table 3 Network architecture of the visual model.

Layer index	Operation
0	122 × 122 × 29 Input
1	3×3 conv, 96, ReLU, 3×3 max-pool
2	3×3 conv, 256, ReLU, 3×3 max-pool
3	3×3 conv, 512, ReLU
4	3×3 conv, 512, ReLU
5	3×3 conv, 512, ReLU, 3×3 max-pool
6	4,096 dense, ReLU
7	4,096 dense, ReLU
8	500 dense, softmax

Table 4 Word recognition accuracy [%] of the audio-only and visual-only models.

	Cond. 1	Cond. 2	Average
Audio-only	94.72	92.35	93.54
Visual-only	49.11	40.61	44.86

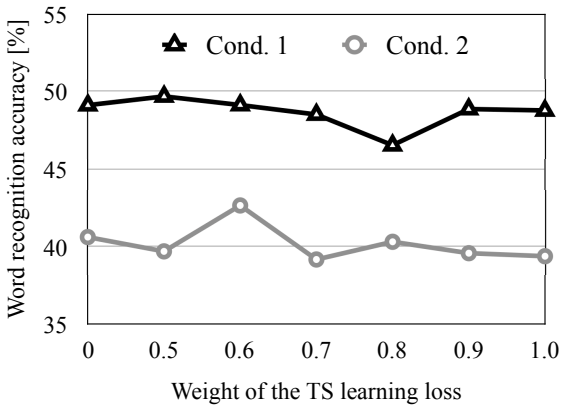

Fig. 2 Word recognition accuracy [%] of lip reading using the TS learning.

が性能向上率が大きい理由として，TS 学習によるモデルの正則化が考えられる．厳密な正解ラベルだけでなく，厳密には正解ではないが学習しやすい教師モデルの出力を用いて生徒モデルを学習することは，モデルの正則化の効果があると考えられ，実験によりこれを確認した．

6 おわりに

本研究では，teacher-student (TS) 学習を用いた，音声認識ネットワークからリップリーディングネットワークへの知識の蒸留を提案した．TS 学習は，主にモデル圧縮に応用され，大規模な教師モデルから小規模な生徒モデルを学習するために利用される．本稿では，TS 学習を異なるモダリティ間での知識蒸留に応用し，情報量の多い音声認識モデルを用いて，情報量の少ないリップリーディングモデルを学習した．評価実験により，リップリーディングにおいて，音声認識モデルの出力を用いた TS 学習の有効性を示した．また，学習データ量が少ない場合に，TS 学習が正則化として働き，より大きく性能を向上させることを確認した．本稿では単語単位の認識を行なったが，今後は文単位の認識に対する TS 学習の評価を行う．

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 *ASRU*, vol. 1, pp. 71–74, 1999.

[3] K. Noda *et al.*, “Audio-visual speech recognition using deep learning,” Applied Intelligence, vol. 42, no. 4, pp. 722–737, 2015.

[4] A. H. Abdelaziz, “NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition,” in *INTERSPEECH*, pp. 3752–3756, 2017.

[5] G. Hinton *et al.*, “Distilling the knowledge in a neural network,” in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[6] Y. Chebotar and A. Waters, “Distilling knowledge from ensembles of neural networks for speech recognition,” in *INTERSPEECH*, pp. 3439–3443, 2016.

[7] J. Li *et al.*, “Developing far-field speaker system via teacher-student learning,” in *ICASSP*, pp. 5699–5703, 2018.

[8] S. Gupta *et al.*, “Cross modal distillation for supervision transfer,” in *CVPR*, pp. 2827–2836, 2016.

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

[10] J. S. Chung and A. Zisserman, “Lip reading in the wild,” in *ACCV*, pp. 87–103, 2016.

[11] M. Cooke *et al.*, “An audio-visual corpus for speech perception and automatic speech recognition,” The Journal of the Acoustical Society of America, vol. 120, no. 5, pp. 2421–2424, 2006.

[12] E. K. Patterson *et al.*, “Cuave: A new audio-visual database for multimodal human-computer interface research,” in *ICASSP*, pp. 2017–2020, 2002.

[13] B. Lee *et al.*, “Avicar: audio-visual speech corpus in a car environment,” in *INTERSPEECH*, pp. 2489–2492, 2004.

[14] J. S. Chung *et al.*, “Lip reading sentences in the wild,” in *CVPR*, pp. 3444–3453, 2017.