1R-03

口パク動画の発話内容推測における距離学習に基づく精度向上手法

柏木 爽良 † 田中 啓太郎 † 森島 繁生 ‡
† 早稲田大学 ‡ 早稲田大学理工学術院総合研究所

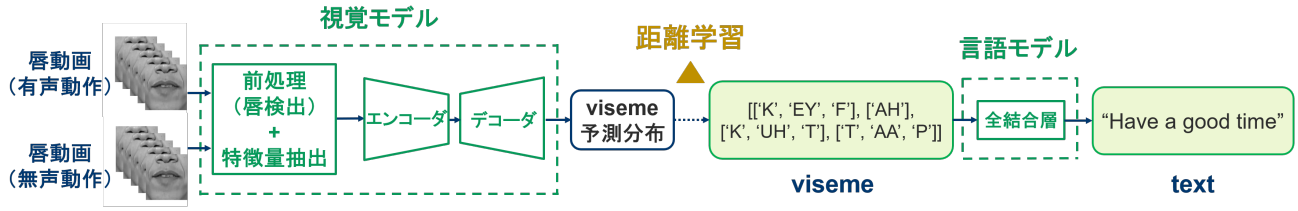

図 1: 距離学習を用いた提案手法の概要図

1. はじめに

唇の動きのみから発話内容を推測するタスクを visual speech recognition (VSR) [1] と呼ぶ．VSR は読唇術としても知られ，声を出すのが困難な人のコミュニケーション円滑化や，声を出せない公共の場におけるスマートフォンの操作に有用である．既に実用化されている例もあり，代表的なものとして音声を用いずに唇の動きによってテキストの入力を行うシステム，silent speech interfaces (SSI) [2] が挙げられる．

一般に唇の動きは，声を出す通常の話し方（有声動作）と声を出さない口パクの話し方（無声動作）とで異なる．しかし，既存のデータセットの多くは有声動作で構成されており，SSI のように無声動作の発話内容を推測する場合も，有声動作のデータセットで訓練されたモデルが使用される．このため，実使用時の動作である無声動作に対する予測精度が，有声動作に対する予測精度を大きく下回るという問題が生じる [3].

本研究では，唇の動きの最小単位を表す viseme に着目することで，有声動作と無声動作の唇の動きの違いを吸収し，同等精度での予測を目指す．有声動作と無声動作で実際の唇の動きは異なるが，viseme 系列は発話された文章のみに依存するため，共通の viseme で表される．そこで，有声動作・無声動作間の同一 viseme ラベルに対応するモデルの出力に対して，距離学習を行う枠組みを提案する．加えて，訓練時に用いる無声動作のデータを減らすことで，有声動作に比べて無声動作のデータが少ない現状を再現する．評価実験を通して，viseme に着目した距離学習の有用性と，提案手法によって限られた無声動作のデータを効果的に活用できることを示す．

2. 関連研究

2.1 有声動作と無声動作の違い

人間は話す時，無意識のうちに自分自身の音声をフィードバックとして利用しながら目的とする音声を発している．しかし，無声動作では音声によるフィードバックを得ることができないため，代わりに体性感覚によるフィードバックを強く意識するようになり，有声動作の唇の動きとの違いが生じる．顔表面に取り付けた電極から発話時の筋肉の動きの大きさを測定し，有声動作と無声動作の唇の動きを比較した研究 [4] では，無声動作の方が唇を閉じる音（“b”，“p”，“m” 等の子音）で大きな電位が得られ，唇の動きは大きくなる傾向が示されている．

2.2 Viseme

聴覚上の音声の最小単位を表す phoneme に対し，同じ唇の動きを示す phoneme をまとめた視覚上の音声の最小単位を viseme と呼ぶ. 従来 VSR の研究では，文字や単語をクラスとして唇動画から直接発話文章を推測する End-to-end モデルが使用されてきた．それに対し Fenghour ら [5] は，唇動画から viseme 系列を推測する視覚モデルと，viseme 系列から発話文章を推測する言語モデルから成る，多段階モデルを提案した. 多段階モデルでは，文字や単語を扱う場合に比べて少ない数の viseme をクラスとして用いるため，視覚モデルにおいて前後の文脈を考慮する必要がない. このため，訓練データに含まれない未知の語彙に対しても同一のモデルでの viseme 予測が可能である．また，有声動作と無声動作の唇の動きの対応関係も，視覚モデルにおいて viseme の種類数のみ学習すればよい．

3. 提案手法

本研究では，Fenghour ら [5] の viseme を介する多段階モデルを採用する．有声動作と無声動作の同一の viseme ラベルに対応する viseme 予測分布を近づけるよう，視覚モデルの出力に対して距離学習を行う（図 1）．視覚モデルの損失関数 $\mathcal{L}_{\text{total}}$ には，viseme の分類を行う交差エントロピー誤差 $\mathcal{L}_{\text{CE}}$ に加え，有声動作と無声動作の違いを吸収する JS ダイバージェンス $\mathcal{L}_{\text{JS}}$ を使用する．

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{JS}} \tag{1}$$

ただし，$\mathcal{L}_{\text{JS}}$ は式 (2)-(3) のように表される．

$$\mathcal{L}_{\text{JS}} = \mathcal{D}_{\text{JS}}(P||Q) \tag{2}$$

$$\mathcal{D}_{\text{JS}}(P||Q) = \frac{1}{2}\mathcal{D}_{\text{KL}}(P||M) + \frac{1}{2}\mathcal{D}_{\text{KL}}(Q||M) \tag{3}$$

ここで，有声動作における各 viseme ラベルに対応する viseme 予測分布の平均を，各 viseme の正解分布とする．確率分布 P には，正解ラベルの viseme 系列に対し，系列を構成する viseme の正解分布をそれぞれ対応させたものを用いる．また，確率分布 Q は，視覚モデルの出力である viseme 予測分布を表す．すなわち，確率分布 P は確率分布 Q の正解分布となる．したがって $\mathcal{L}_{\text{JS}}$ は，有声動作と無声動作の viseme 予測分布間の距離を小さくする効果に加え，有声動作・無声動作に関わらず同一の viseme ラベルに対応する全ての viseme 予測分布同士

Visual Speech Recognition for Silent Speech using Metric Learning:
Sara Kashiwagi†, Keitaro Tanaka†, and Shigeo Morishima‡ (†Waseda University, ‡Waseda Research Institute for Science and Engineering)

[pad], AA, AH, AO, CH, ER, EY, F, IY, K, P, T, UH, W, <sos>, <eos>, [space]

図 2: viseme 系列に含まれる viseme と特殊記号

を近づける効果を有する．確率分布 M と KL ダイバージェンス $\mathcal{D}_{\mathrm{KL}}(P||Q)$ は，式 (4)-(5) のように定義される．

$$M = \frac{1}{2}P + \frac{1}{2}Q \tag{4}$$

$$\mathcal{D}_{\mathrm{KL}}(P||Q) = P \log \frac{P}{Q} \tag{5}$$

4. 評価実験

4.1 データセット

39 人の話者による有声動作・無声動作での各 50 回の発話で構成される AV Digits Database [3] を使用した．20 人を訓練データ，8 人を検証データ，11 人をテストデータとして用い，話者に依存しない条件で実験を行った．語彙は，10 種類の簡単な英語の文章（“Excuse me”，“Goodbye”，“Hello”，“How are you”，“Nice to meet you”，“See you”，“I am sorry”，“Thank you”，“Have a good time”，“You are welcome”）から成る．The Carnegie Mellon Pronouncing (CMU) Dictionary を用いて発話文章を phoneme 系列に変換し，さらに Lee らの分類 [6] を用いて phoneme 系列を viseme 系列に変換することで，唇動画とそのラベルとなる viseme 系列のデータセットを作成した．なお，viseme 系列は図 2 のように，13 種類の viseme と 4 種類の特殊記号 (系列の長さを揃える “<pad>”，系列の始まりを示す “<sos>”，系列の終わりを示す “<eos>”，単語の区切りを示す “[space]”) で表現される．

4.2 実験条件

まず，学習データの話し方の種類による精度の違いを評価するため，有声動作のみで訓練を行い，有声動作に対する精度と無声動作に対する精度を比較した．続いて，有声動作と無声動作の両方を用いて訓練を行い，提案手法である距離学習を導入した．以上の実験では有声動作・無声動作で各 1000 個の訓練データを用いた．さらに無声動作のデータが少ない状況における距離学習の効果を評価するため，訓練時に用いる無声動作のデータを 20% 刻みで減らした時の無声動作に対する精度を比較した．具体的には，話者一人当たりのデータ数を 50 個から 10 個ずつ減らし，データを変えて 5 回ずつ実験を行った．

評価指標は viseme error rate (VER) と sentence accuracy rate (SAR) を用いた．VER は視覚モデルの出力における viseme の誤り率を表し，次式 (6) で計算される．

$$\mathrm{VER} = \frac{V_S + V_D + V_I}{V_N} \tag{6}$$

ただし，V_N は viseme の総数，V_S は誤った viseme に置換された viseme の総数，V_D は誤って削除された viseme の総数，V_I は誤って挿入された viseme の総数を意味する．また SAR は，視覚モデルで予測された viseme 系列を言語モデルで 10 個の語彙に分類した際の正解率を表す．

4.3 実験結果

評価実験の結果を表 1 に示す．有声動作のみで訓練を行い，無声動作でテストを行った場合，有声動作でテストを行った場合に対して VER は 4.87%増加，SAR は 4.55% 減少し，訓練データとテストデータの話し方の種類の不一致による精度低下が確認された．次に，提案手法である距離学習を導入すると，距離学習を用いない場合に対して VER は 0.20%減少，SAR は 1.09%向上し，提案手法の有効性が確認された．さらに，訓練時に用いる無声動作のデータ数を 1000 個から 800 個，600 個，400 個，200 個と減らし，無声動作に対する VER を比較した結果を図 3 に示す．ただし，縦軸は VER，横軸は無声動作の訓練データ数を表し，VER の平均を m，標準偏差を σ として，m を実線で示し，$m - \sigma$ から $m + \sigma$ の領域を色付けしている．距離学習ありの VER が距離学習なしの VER を常に下回り，さらに無声動作のデータを減らした時の VER 増加を抑制できていることがわかる．

表 1: 評価実験の精度比較

距離学習	訓練データ	テストデータ	VER (%) ↓	SAR (%) ↑
なし	有声	有声	4.87	96.55
なし	有声	無声	9.74	92.00
なし	有声・無声	無声	6.18	95.64
あり	**有声・無声**	**無声**	**5.98**	**96.73**

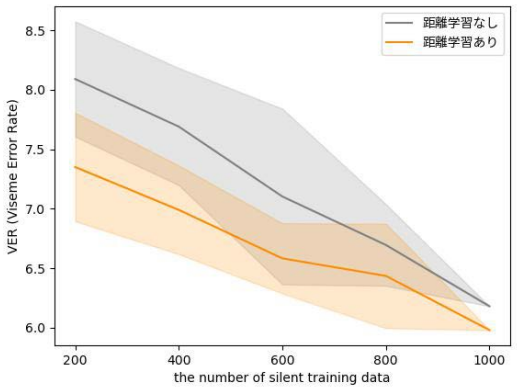

図 3: 無声動作の訓練データ数減少に伴う精度推移

5. おわりに

本稿では，無声動作に対する発話内容予測精度向上を目指し，viseme を介した距離学習に基づく手法を提案した．評価実験の結果，距離学習によって精度が向上し，加えて，訓練時に用いる無声動作のデータを減らした場合も本手法の有効性が確認された．本稿では 10 個の簡単な語彙に限定したが，今後は，膨大な有声動作のデータセットと End-to-end モデルで距離学習を行うことで，10 個の語彙に含まれない未知の語彙の無声動作に対して有声動作と同等の精度で VSR を行うことを目指す．

謝辞 本研究は，JSPS 科研費 (19H04137，21H05054，22J22424) の補助を受けています．

参考文献

[1] P. Ma *et al.*: “Visual speech recognition for multiple languages in the wild,” *Nature Machine Intelligence*, 1–10, 2022.

[2] L. Pandey *et al.*: “LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model,” *CHI Conference on Human Factors in Computing Systems*, 1–19, 2021.

[3] S. Petridis *et al.*: “Visual-only recognition of normal, whispered and silent speech,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6219–6223, 2018.

[4] Y. Janke *et al.*: “Impact of lack of acoustic feedback in EMG-based silent speech recognition,” *Interspeech*, 2010.

[5] J.S. Fenghour *et al.*: “Lip Reading Sentences Using Deep Learning With Only Visual Cues,” *IEEE Access*, 8:215516—215530, 2020.

[6] S. Lee *et al.*: “Audio-to-visual conversion using Hidden Markov models,” *Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence*, 563–570, 2002.