

マルチモーダル音声認識における 音声と画像の協調によるモデル適応法の検討

Model Adaptation Using Audio-Visual Interactivity for Multi-Modal Speech Recognition

絹田 卓也†
Takuya Kinuta

田村 哲嗣‡
Satoshi Tamura

速水 悟‡
Satoru Hayamizu

1. はじめに

近年音声認識技術の発達により、携帯電話での検索機能、カーナビゲーションシステムの音声操作などで、音声認識技術が利用されている。音声による操作や情報検索の最大の利点は、ユーザーのリテラシーに関係なく、誰でも簡単にそれらの機器を扱うことができる点にある。しかし、実環境では、周囲の雑音が大きな場所で音声認識が利用されることが多い。そこで、音響情報と画像情報を用いたマルチモーダル音声認識による、雑音環境に頑健な音声認識が研究されている[1,2]。

一方、音声認識結果と、認識対象話者の少量の発話データを用いてモデル適応を行い、認識結果を改善する手法がある[3]。同じ発話内容を認識しても、発話者によって認識精度が大きく変わってしまうことがあり、原因として、発話者特有の話す際の癖が挙げられる。この場合、数十名の発話データを用いて学習した。平均的な発話の仕方と比べて大きく特徴が異なってしまう、うまく認識されなくなってしまう。そこで、学習モデルに対し個人適応をかけることでその発話者のクセ、周囲の雑音を学習し、モデルを対象発話者やその環境に合わせることで認識精度を上げることを考える。大西らは、マルチモーダル音声認識を行う際に、適応したモデルの音響・画像情報の重みを調節することで、一方の認識精度が低くても、他のモダリティに重みを増やすことによって認識精度が向上することを確認している[4,5]。しかしながら、各モダリティの認識精度が低い場合、適応を行っても十分な認識精度にならないことが課題となっている。

そこで、本研究では音響モデルまたは画像モデルに対し、音声のみ、画像のみのユニモーダルの認識結果よりも認識精度の高い、マルチモーダルの認識結果を適応に用い、各モダリティの認識精度を向上させる。これにより、先述した問題を解決し、マルチモーダル音声認識の精度向上を目指す。また、適応後の精度の上がった認識結果を用い、各モダリティに繰り返し適応させることで、より発話者に適応したモデルを作り、更なる認識精度の向上を目指す。なお、実環境下において、適応に用いる正解ラベルを人手によって、すべて作成するのは実用的でないことから、本研究では教師なしによる適応を行う。

2. マルチモーダル音声認識

マルチモーダル音声認識の流れを図1に示す。マルチモーダル音声認識において、音声と画像という異なるモダリティから得られる情報を同一に扱うよりも、状況に応じてそれぞれの情報の重みを変えて利用した方が有効であると考えられる。そこで、初期統合法によるマルチモーダル音声認識では、各モダリティの情報をストリームという形で纏め、新たなパラメータによって各々のストリームの重みを制御することができるマルチストリームHMMを用いる。本研究ではHMMの状態数8、混合数を音声36、画像16で扱う。このマルチストリームHMMでは、遷移後の状態が q_j となったときに特徴量 o_t を観測する出力確率に混合ガウス分布を用いている。出力確率 $b_j(o_t)$ は式(1)のように表すことができる。

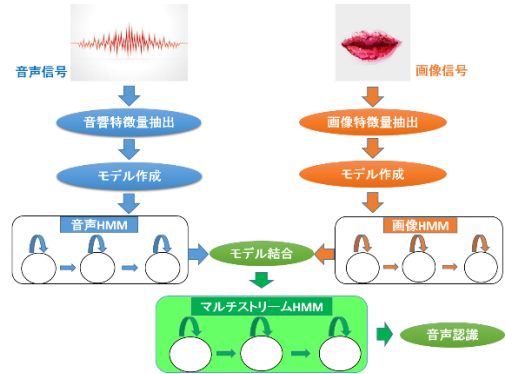


図1 マルチモーダル音声認識の流れ

$$b_j(o_t) = \prod_{s=1}^S \left\{ \sum_{m=1}^{M_{s,j}} c_{s,j,m} N(o_{s,t}; \mu_{s,j,m}, \Sigma_{s,j,m}) \right\}^{\lambda_s} \quad (1)$$

ここで S はストリームの個数、 $M_{s,j}$ はストリーム s の状態 q_j における正規分布の数、 $c_{s,j,m}$ は m 番目の正規分布の混合重み、 $o_{s,t}$ はストリーム s の特徴量ベクトル、 λ_s はストリーム重みである。また、 $N(o; \mu, \Sigma)$ は特徴量 o の正規分布で式(2)のように表せる。

$$N(o; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(o - \mu)^T \Sigma^{-1}(o - \mu)\right) \quad (2)$$

ここで n は o の次元数、 μ は平均ベクトル、 Σ は共分散行列である。

雑音などの影響で信頼度が低下したストリームがあれば、そのストリームの重み λ_s を低く設定することで、より信頼度の高いモダリティを用いた認識が可能となる[6]。本研究ではストリーム重みを手動で設定して実験を行い、認識精度が一番高い時のストリーム重みを採用している。

3. モデル適応

モデル適応には、音声認識で代表的な手法であるMLLR(Maximum Likelihood Linear Regression)法[7]を用いる。MLLR法は回帰行列を少量の適応発話から推定し、回帰行列によってモデルパラメータを線形変換する手法で、適応するモデルの正規分布の平均ベクトルを推定する。式(3)で表されるように、HMMの正規分布の平均ベクトル μ の線形変換によって適応後の平均ベクトル $\hat{\mu}$ が与えられる。

$$\hat{\mu} = H\mu + b \quad (3)$$

H は n 次元正方行列であり、 b は n 次元のバイアスペクトルである。これらは適応データをもとに推定している。

4. 提案手法

4.1 音声認識・モデル適応

はじめに音響および画像モデルを学習、結合し、マルチモーダル音声認識を行う。ここで作成されたモデルによるマルチモーダル音声認識結果を適応ラベルとして、音響および画像モデルを適応し、新たなモデルを結合する。これを用いてマルチモーダル音声認識を行う。さらにこの認識結果を適応に用いることで適応を複数回行い、より認識対象話者に適応した音声認識を行う。なお、本研究ではテストデータの認識結果を用いてclosed適応とした。

† 岐阜大学大学院 工学研究科

‡ 岐阜大学 工学部

提案手法の流れを以下に示す.

- ① 音響および画像特徴量抽出を行う.
 - ② 音響および画像モデルを作成し, 統合する.
 - ③ マルチモーダル音声認識を行う.
 - ④ ③の認識結果を用いて, 音響および画像モデルに対し, 個人適応を行う.
 - ⑤ 音響・画像モデルの結合を行う.
 - ⑥ ③・④・⑤を繰り返し, 複数回適応を行う.
- また, 上記の流れを図 2 に示す.

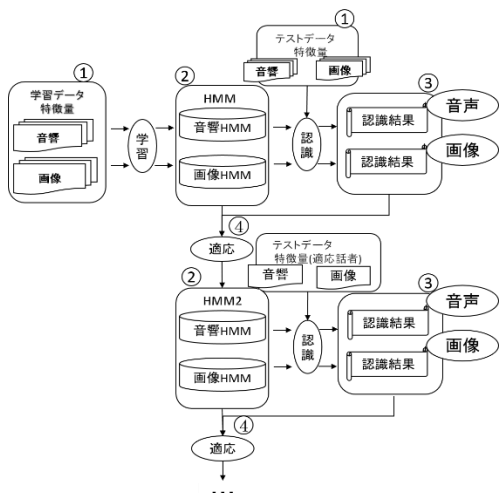


図 2 提案手法の流れ

4.2 音声特徴量

本研究では, 音声特徴量にメル周波数ケプストラム係数 (MFCC) を用いた. MFCC は, 音響特徴量として話者識別や音声認識で広く用いられている. 周波数領域に変換したスペクトル情報に対し, ヒトの周波数知覚特性上重要な成分を考慮した重みづけをした特徴量である.

4.3 画像特徴量

画像特徴量には固有顔を用いた. 固有顔は CENSREC-1-AV[8]のベースラインによって, 画像に対しサンプリング, サイズ縮小, グレースケール化, ラスタスキャンを掛けてベクトル化, ベクトルの正規化を行う. その後主成分分析を行い, 固有値を求め, 固有値の大きいもの 10 個(次元数)を特徴量としたものである.

5. 実験

本研究では, 提案手法による精度の向上を確認するため, 適応にユニモーダルの認識結果を用いて実験した実験 A と, 4 章で説明した提案手法である, マルチモーダルの認識結果を適応に用いた実験 B に分けて実験を行った.

5.1 実験条件

実験には CENSREC-1-AV のデータを用いて行った. 実験環境を表 1 に示す. なお, モデル適応は話者ごとに, 全てのテストデータを用いて行った. 発話数字の読みを表 2 に示す.

評価は認識精度 Acc (Accuracy)で行った. これは式(4)で表される.

$$Accuracy(\%) = \frac{H - I}{N} \times 100 \quad (4)$$

ただし, H : 正解数, I : 挿入誤り, N : ラベルの総数である.

表 1 実験環境

	男性	女性
テストデータ	25 名	26 名
	1963 発話(91 分)	
学習データ	22 名	20 名
	3234 発話(149 分)	
特徴量	音響	MFCC12 次元, パワー1 次元, Δ , $\Delta\Delta$ (計 39 次元)
	noise	clean, ホワイトノイズ SNR0dB, 10dB, 20dB
	画像	固有顔 10 次元, Δ , $\Delta\Delta$ (計 30 次元)
	noise	clean, ガンマ補正
発話内容	日本語連続数字 1~7 桁	

表 2 数字の読み

単語	読み
1(one)	/ichi/
2(two)	/ni/
3(three)	/saN/
4(four)	/yoN/
5(five)	/go/
6(six)	/roku/
7(seven)	/nana/
8(eight)	/hachi/
9(nine)	/kyuH/
Z(zero)	/zero/
0(oh)	/maru/

5.2 実験 A ユニモーダル教師なし適応

実験 A の流れを以下に示す.

- ① 音響および画像特徴量抽出を行う.
- ② 音響および画像モデル作成を行う.
- ③ 音声単体・画像単体の音声認識を行う.
- ④ ③の認識結果を用いて, モデルに対し個人適応を行う.
- ⑤ ②・③・④を繰り返し行う.

また, 上記の流れを図 3 に示す.

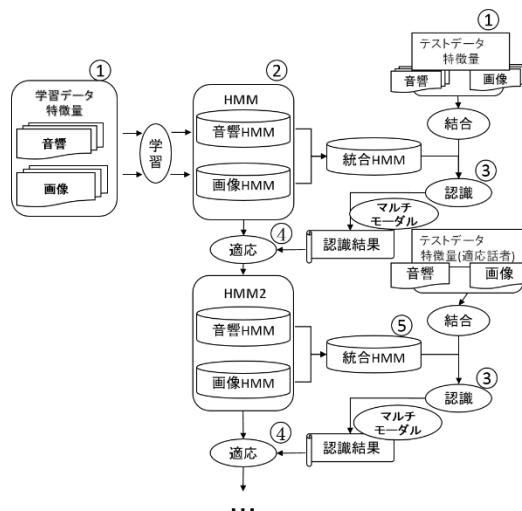


図 3 実験 A の流れ

5.2.1 実験結果 A

音声単体, 画像単体での音声認識結果をそれぞれ A, V, マルチモーダルでの認識結果を AV と表す. 本研究では, 適応を繰り返し行うため, 繰り返し回数によって結果を表 3 のように表記する.

表3 結果表記方法

適応回数	0	1	2
音声	A	A2	A3
画像	V	V2	V3
マルチモーダル	AV	AV2	AV3

実験 A は音声にホワイト雑音を 3 種類の SNR で重畳したものを、画像にノイズとしてガンマ補正をかけたものおよび、クリーン画像を用いてそれぞれ行った。結果を図 4 に示す。グラフの縦軸は認識精度である。

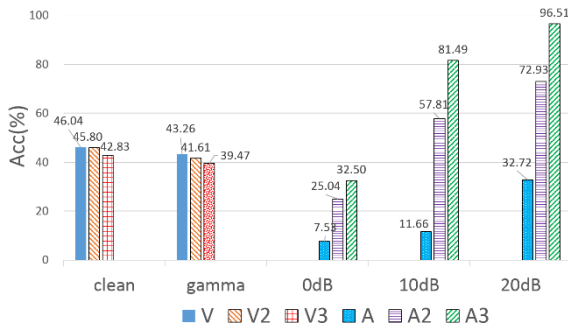


図4 実験 A 結果

5.2.2 考察

図 4 より、音声雑音 0dB の A, A2, A3 から、適応無音声で 7.5% 程度の認識率が、2 度適応を行うことで 32.5% に向上、雑音 20dB では 32.7% 程度の認識精度が、96.5% にまで向上している。このことから、音声単体の認識結果を用いた適応は、雑音が大きき場合でも有効であることが確認できる。一方、図 4 の V, V2, V3 より、クリーン画像、ガンマ画像単体での適応認識結果は、適応を繰り返してもほとんど変化しなかった。適応前の認識精度は音声情報よりも画像情報の方が高く、その認識結果を適応に用いているにも関わらず画像情報の精度が向上しないことから、MLLR 法によるモデル適応は、画像情報より音声情報に対し有効であることが考えられる。そして、精度の変化がみられない原因として、画像情報に正しくモデル適応がされていなかったことが考えられる。

認識率が向上した話者の多くは、元から認識率が高い話者であり、精度の低い話者の改善につながらなかった。このことから、精度の低いモダリティを改善するためには、精度の良い認識結果が必要と考えられる。しかしながら、図 4 のとおり、画像モダリティでは精度の高い認識結果を得ることは難しい。そこで、他の精度の高いモダリティで認識した認識結果を適応に用いるで、精度の改善を検討する必要がある。

5.3 実験 B マルチモーダル教師なし適応

ここでは、4 章で説明した提案手法での実験を行う。実験 A との主な違いは、データは実験 A と同じものに加え、クリーン音声を使用したこと、適応にマルチモーダル音声認識結果を用いたことである。

5.2.1 実験結果 B

実験 B は以下の 3 つの条件で実験を行った。

- ・条件 1: 音声雑音+クリーン画像
- ・条件 2: 音声雑音+ガンマ画像
- ・条件 3: クリーン音声+ガンマ画像

それぞれ結果を図 5, 図 6, 図 7 に示す。

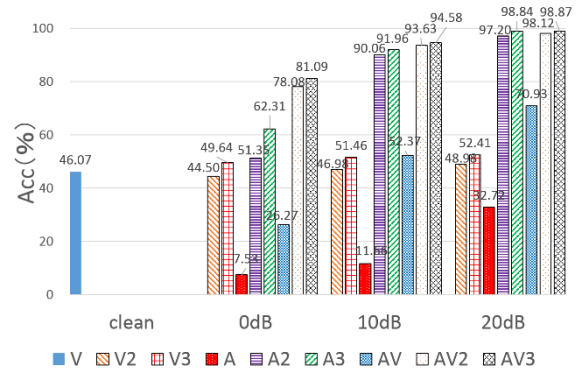


図5 条件 1 (音声雑音+クリーン画像)

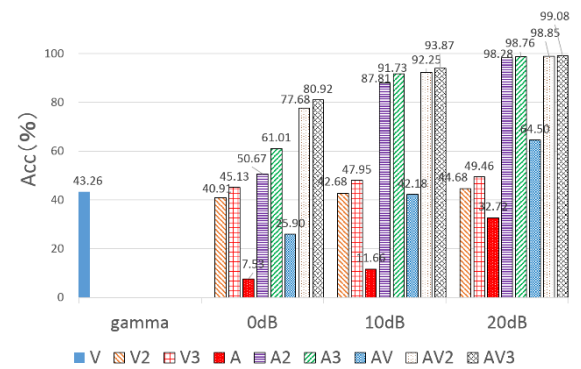


図6 条件 2 (音声雑音+ガンマ画像)

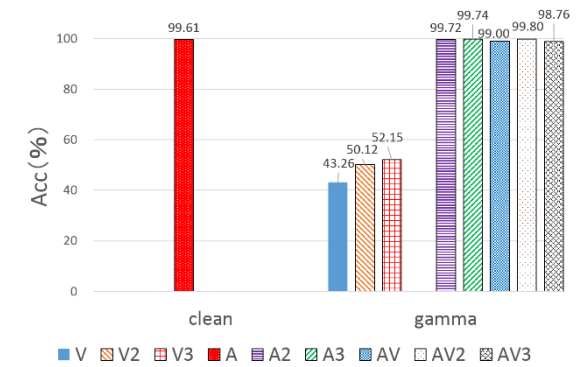


図7 条件 3 (クリーン音声+ガンマ画像)

5.2.2 考察

図 5, 6 より、音声、画像での適応による認識精度の向上が確認できる。また、条件 1 の画像にガンマ補正をかけた条件 2 では、画像とマルチモーダルの認識精度が低下してはいるものの、音声とマルチモーダルの認識結果はほぼ同程度まで向上している。これは、精度の低いモダリティである画像情報よりも、精度の高いモダリティである音声情報の与える影響が大きいためと考えられる。一方、図 7 では、音声の精度がほぼ 100% のモダリティと、適応によって精度の向上した画像情報を統合したにも関わらず、わずかながらマルチモーダルの精度が低下している。これは音声認識の精度が非常に高いので、画像のモダリティが悪影響を及ぼしていることが考えられる。

図 8 に、音声雑音 0dB における実験 A と、条件 2 における実験 B の、A3 での話者別の結果を示す。実験 A と比較して、実験 B では適応を 2 度行った雑音 0dB での、音声単体の認識結果が 30% 近く向上している。これにより、マルチモーダル音声認識結果を適応に用いることによって、音声のみの認識においても、すべての発話者の認識精度の大幅な向上を確認できた。

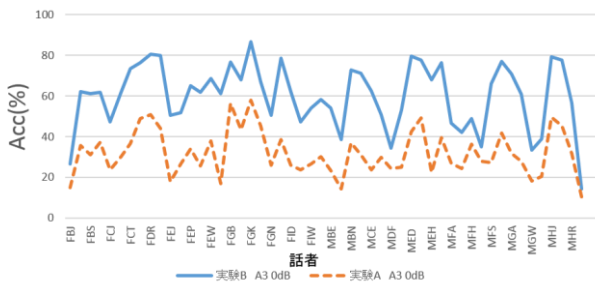


図8 音声認識結果 (雑音 0dB, 実験A—実験B条件2)

図9に、条件2での適応の前後における話者別の画像認識結果の変化を示す。先述したように実験Aでは、適応を繰り返すたび画像の認識精度が低下したが、実験Bでは適応を繰り返すことで43.3%から45.1%になり、若干ながら精度の向上を確認できた。ただし、こちらは音声の場合と異なり、平均認識精度は向上したが、適応をすることで、認識精度が低下する話者も確認された。

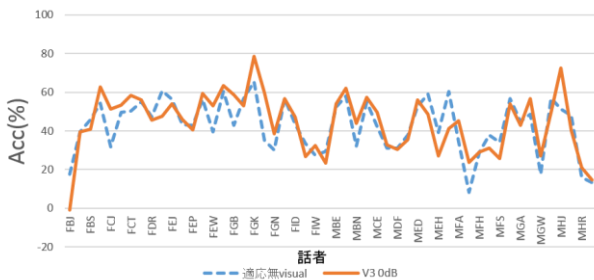


図9 画像認識結果 (実験B条件2, 適応前—適応後)

図10に、条件2での適応後の音声・画像・マルチモーダルでの認識結果を話者別に並べたものを示す。音声・画像それぞれ適応したモデルを統合させて認識を行ったマルチモーダル音声認識では、全体的に音声単体・画像単体での認識結果より大幅に精度が向上している。雑音0dBでの環境下では、比較的精度のよい音声認識単体での精度より20%程精度が向上しており、雑音環境下でも高い認識精度が確認できた。また実験Aでの適応後の音声での認識精度が37%、画像での認識精度が60%の話者の場合、実験Bでこの画像情報を含むマルチモーダル情報で適応を行った結果、音声単体での精度が72%まで向上した。このことから、必ずしも音声情報が画像情報より精度向上に貢献するわけではなく、音声・画像の精度の高い認識結果が認識精度向上に役立っているのではないかと考えられる。

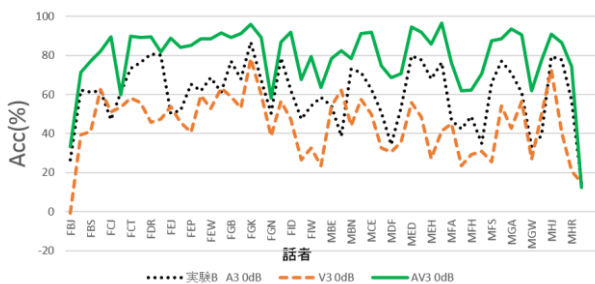


図10 モダリティ比較 (実験B条件2, 雑音0dB)

6. まとめ

本研究では、マルチモーダル音声認識における、マルチモーダル認識結果を適応データに用いる個人適応の有効性を検討した。提案手法と適応データにユニモーダル音声認識結果を用いた実験を比較して、認識結果の向上が見られたことから、適応データにマルチモーダル音声認識結果を用いた提案手法の有効性を確認した。今後の課題として、本研究では手動で行っている最適なストリーム重みの設定を、

実環境を想定した、自動で設定を行う方法の検討が必要となる。

参考文献

- [1] 高山俊輔, 松尾俊秀, 岩野公司 “対話システムへの利用を想定したマルチモーダル音声認識の検討” 電子情報通信学会技術研究報告. SP, 音声 vol.107, no.77, pp.19-24, 2007.
- [2] 朽木陽佑 “口唇動画像を用いた区分的線形変換による雑音環境下マルチモーダル音声認識” 日本音響学会春季講演論文集. pp.25-28, 2012.
- [3] 中村哲 “音声認識における話者適応” 電子情報通信学会技術研究報告. SP, 音声 vol.94, no.42, pp.17-24, 1994.
- [4] 大西正真, 田村哲嗣, 速水悟 “音声・画像のモダリティ間の相互作用に着目した音声認識のモデル適応” 電子情報通信学会技術研究報告. SP, 音声 vol.111, no.97, pp.17-20, 2011.
- [5] 大西正真, 田村哲嗣, 速水悟 “マルチモーダル音声認識のモデル適応における音響・画像の相互影響” 日本音響学会春季講演論文集. 2-P-22, pp.179-180, 2011.
- [6] 田村哲嗣, 岩野公司, 古井貞熙 “尤度比最大基準によるストリーム重み最適化を用いたマルチモーダル音声認識の性能評価” 日本音響学会春季講演論文集. 3-8-1, pp.123-124, 2004.
- [7] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous-density hidden Markov models,” Computer Speech and Language, vol.9, pp.171-185, 1995.
- [8] S. Tamura et al., “CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition,” Proc.AVSP2010, pp.85-88, 2010.