

初期統合によるバイモーダル大語彙連続音声認識

Audio-visual large vocabulary continuous speech recognition based on feature integration

石川 剛[†] 澤田 裕子[†] 全 炳河[†] 南角 吉彦[†]

Tsuyoshi ISHIKAWA Yuko SAWADA Heiga ZEN Yoshihiko NANKAKU

宮島 千代美[†] 徳田 恵一[†] 北村 正[†]

Chiyomi MIYAJIMA Keiichi TOKUDA Tadashi KITAMURA

1. まえがき

音声認識では、周囲の雑音や環境の違いにより認識率が低下する問題がある。雑音や環境の違いによる影響を緩和する方法の一つとして、音声情報に唇動画像情報を併用して認識を行うバイモーダル音声認識が挙げられる。これまでバイモーダル音声認識に関するさまざまな研究が行われ、その有効性が示されている。しかし、大語彙連続音声認識に関する研究は英語音声の研究 [1] がある程度というのが現状であり、日本語音声についての研究はまだ報告されていない。そこで本研究では、日本語音声のバイモーダル大語彙連続音声認識について初期統合を用いた実験を行い、その有効性について検討する。

2. 主成分分析 (PCA) による画像の特徴抽出

バイモーダル音声認識において、音声の特徴量はメルケプストラムなどのある程度定まった特徴量が用いられるのに対し、画像の特徴量はその特徴抽出の違いから、画像ベース法やモデルベース法などさまざまな手法が提案されている。本研究では、画像ベース法に基づいた PCA による特徴量を用いる。

1 枚の画像の全画素値を並べた M 次元ベクトルを $x = [x_1, x_2, \dots, x_M]^T$ とする。 N 枚の画像 x_1, x_2, \dots, x_N を用意し、画像 x から、平均画像 \bar{x} を減算したものを $\hat{x} = x - \bar{x}$ と表す。このとき、 N 枚の画像から得られる行列 $\hat{X} = [\hat{x}_1, \hat{x}_2, \dots, \hat{x}_N]$ に対して PCA を行うことによって正規直交基底 $U = [u_1, u_2, \dots, u_N]$ が得られる。これは固有ベクトルと呼ばれ、各固有ベクトルが唇のさまざまな特徴を表した固有唇である。また、再構成画像 \hat{x} は U の線形結合によって $\hat{x} = Uy$ と表すことができる。ここで、主成分スコア $y = [y_1, y_2, \dots, y_N]^T$ は元の画像 x を表す特徴量と考えることができる。そのため、 y の次元数 N より小さな n 次元を用いることで、特徴空間の次元を圧縮することができる。図 1 に、PCA による画像圧縮の例 ($N = 1000$) を示す。全体的にぼやけた画像であるが、低次元でも十分特徴を表していることがわかる。



原画像 再構成画像 再構成画像 再構成画像
($n=5$) ($n=15$) ($n=25$)

図 1: 原画像と再構成画像

3. 大語彙連続音声認識

大語彙連続音声認識の標準的なアプローチは、復号列が最大事後確率を持つように、音響的な観測系列に基づいて単語列を復号することである。音声時系列パターンとして、 $O = (o_1, o_2, \dots, o_T)$ が入力として与えられたとする。ここで、 o_t は第 t 番目のフレームにおける音声の特徴ベクトルである。このとき、入力音声 O に対する単語列 W は、次式で与えられる。

$$\hat{W} = \arg \max_W P(W|O) \quad (1)$$

また、ベイズの定理により

$$\hat{W} = \arg \max_W P(O|W)P(W) \quad (2)$$

と書くことができる。ここで、 $P(O|W)$ は HMM でモデル化された音響モデルであり、 $P(W)$ は W の事前生起確率で、言語モデルによって与えられる。また、音響モデルと言語モデルの重要度の割合を表した言語重み α と挿入誤りや脱落誤りを制御するための挿入ペナルティ β を導入した次式がよく用いられる。

$$\hat{W} = \arg \max_W \{\log P(O|W) + \alpha \log P(W) + n\beta\} \quad (3)$$

ここで、 n は単語列 W に含まれる単語数である。

4. 音声と画像の統合法

バイモーダル音声認識における音声と画像の統合方法には、時間的な統合単位の違いから、初期統合や結果統合などが提案されている。また、音声と唇画像は部分的に非同期になることが報告されており、同期した部分における音声と画像の相関をうまく利用することが重要となる。このような非同期性を考慮した統合法として合成統合 [1] も提案されている。初期統合は、フレーム単位で特徴量を統合する手法であり、音声と画像の相関を利用しているが、非同期性を表現することはできない。しかし、従来の認識デコーダを用いて容易に実現することができるため、本研究では、まず、初期統合による実験を行うこととした。

図 2 に初期統合法による大語彙連続音声認識システムを示す。初期統合では、音声と画像の時刻 t における観測ベクトルを $o_t = (o_{At}, o_{Vt})$ と表した場合、 o_t の状態 i における出力確率は次式により計算することができる。

$$b_i(o_t|M) = b_{A_i}(o_{At}|M)^{\lambda_A} \times b_{V_i}(o_{Vt}|M)^{\lambda_V} \quad (4)$$

ここで、 $b_{A_i}(o_{At}|M)$ は状態 i において音声の特徴ベクトル o_{At} を出力する確率であり、 $b_{V_i}(o_{Vt}|M)$ は状態 i において画像の特徴ベクトル o_{Vt} を出力する確率である。また、 λ_A, λ_V は音声と画像それぞれのストリーム重みである。ストリーム重みは、状態ごと、あるいはモデルごとに異なる値を用いることも可能であるが、本実験では全モデル・全状態共通の値とした。また、 $\lambda_A + \lambda_V = 1$ とした。

[†]名古屋工業大学 知能情報システム学科,
Dept. of Computer Science, Nagoya Institute of Technology

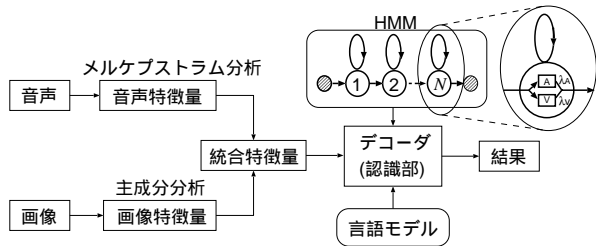


図 2: 初期統合法による認識システム

5. 認識実験

5.1 実験条件

本実験では M2TINIT データベース [2] を使用する。このデータベースは男性話者 1 名の唇動画像とその音声が入録されている。発話内容は ATR 日本語データベースの音韻バランス文 503 文章で用いられているテキストである。学習データには 450 文章を使用し、残りの 53 文章をテストデータとして認識実験を行った。PCA は、学習データからランダムに選んだ 1000 枚の画像に対して行った。音声と画像の分析条件を表 1 に示す。予備実験の結果より、主成分スコア 15 次元を画像の特徴量として用いた。また、音声と画像データのフレーム周期を同期させるため、画像フレームを削除もしくはコピーして補間した。HMM は、状態数 3、混合数 16 の triphone モデルを構築した。言語モデルには、IPA 日本語ディクテーション基本ソフトウェア付属の毎日新聞 45ヶ月分によって作成された 2 万語彙の言語モデル (bigram) を用いた。

5.2 音素認識実験

音響モデルにおける初期統合の効果を確認するため、音声認識実験を行った。テストデータとして clean な音声と SN 比が 24dB, 12dB, 6dB となるように音声中にガウス雑音を加えたデータを用意した。音声の重みを 0.1 単位で変化した場合の認識結果を図 3 に示す。結果より SN 比が高い場合ほど、画像の重みを大きくすることにより、音声のみに比べ認識率が改善されることがわかる。また、clean な音声においても画像の情報を併用することにより、わずかではあるが改善が得られた。

5.3 大語彙連続音声認識実験

次に、言語モデルを用いて大語彙連続音声認識実験を行った。結果を図 4 に示す。音声の重みは、各 SN 比に対して音素認識において最も良い結果となった値で固定した。また、言語重みと挿入ペナルティについては、各 SN 比ごとに設定した。結果より、ノイズが加わることによって音声のみでは認識率が大きく低下することがわかる。特に、24dB から 12dB にかけて、音素認識の結果を比較しても認識率の低下が著しいことがわかる。しかし、唇動画像を用いることによって、12dB においても 44.8% の単語正解精度を維持しており、音声のみの場合と比べ、39% の誤り改善率が得られた。また 6dB においては、音声のみではほとんど認識できていないのに対して、バイモーダル情報では 24.5% の単語正解精度が得られた。以上の結果から、日本語における言語モデルを用いた大語彙連続音声認識においても、バイモーダル情報は有効であることが示された。

表 1: 音声と画像の分析条件

音声	分析窓: Blackman 窓, 分析窓長: 25ms 特徴量: メルケプストラム (18 次), Δ , Δ^2 フレーム周期: 10ms
画像	特徴量: 主成分スコア (15 次元), Δ , Δ^2 フレーム周期: 33.3ms 10ms

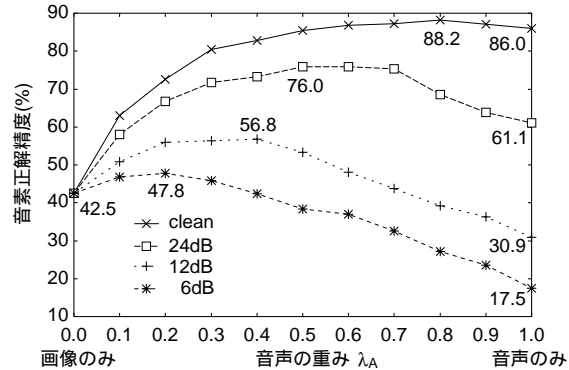


図 3: 音素認識結果

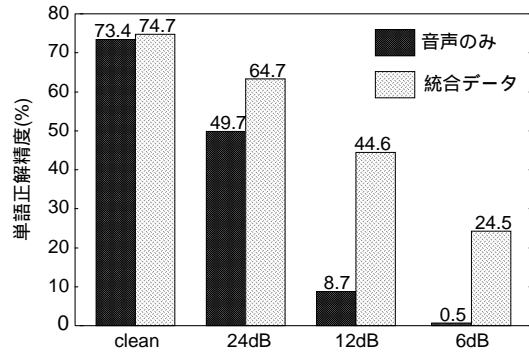


図 4: 大語彙連続音声認識の結果

6. むすび

本研究では、バイモーダル情報を用いた日本語音声の大語彙連続音声認識について、初期統合による認識実験を行った。言語モデルを用いた日本語大語彙音声認識においても、バイモーダル情報の利用によって、音声、画像の単独情報のみより認識率が改善できることを確認した。今後の課題としては、他の統合方法 (結果統合, 合成統合) との比較・検討や、より大規模なデータベースを用いた実験が挙げられる。

謝辞 本研究の一部は、堀情報科学振興財団研究助成、中部電力基礎技術研究所研究助成、および科学研究費補助金若手研究 (B)No.14780274 による。

参考文献

- [1] J. Luetttin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," vol.1, pp.169-172, May 2001.
- [2] 酒向慎司, 近藤重一, 益子貴史, 徳田恵一, 小林隆夫, 北村正, "唇動画像と音声によるマルチモーダルデータベースの構築," 音響学会講演論文集, vol.1, pp.221-223, Mar. 2001.