修士論文

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

2013 年 2 月 6 日

指導教員　広瀬 啓吉 教授

東京大学大学院 情報理工学系研究科
電子情報学専攻　**48116414**

柏木 陽佑

内容梗概

自動音声認識システムは，次世代のコンピュータインタフェースとして非常に注目されている．自動音声認識は 20 世紀半ばから研究が始められ，統計的パターン認識技術の発達や大規模なコーパスの構築とともに精度が向上し，特定の環境下では高い精度で大語彙の連続音声認識を実現出来るようになった．現在では，携帯電話やカーナビに代表される一般製品のインタフェースとして，また議事録の自動作成や企業のコールセンターにおけるデータマイニングなど，幅広い分野に自動音声認識技術が応用されている．

しかし，自動音声認識システムの精度は，背景雑音の重畳により大幅に低下してしまうことが知られている．この問題を解決するために，非常に多くの研究が行われているものの、未だ十分に解決されていない．そのため，現在実用化されている自動音声認識システムは，雑音が小さな環境下や，ヘッドセットマイクの利用を前提としたものが多い．しかしながら，近年急速に普及した携帯端末における入力インタフェースとしての利用を考えると，高雑音環境下での高精度な音声認識精度の実現への期待は大きい．

実環境での自動音声認識の使用を想定した場合，雑音に対して頑健なシステムを構築することは非常に重要である．本研究では，雑音環境下での認識性能の向上のために，音声そのものだけなく，口唇画像情報にも注目する．ヒトが高雑音環境下で音声認識を行うプロセスに注目すると，音声だけでなく，顔の表情や口唇の動きなどによって認識を補助していると考えられる．例えば，口唇の形や動きによってある程度の音素や発話タイミングを把握する経験は誰もが経験しているものであろう．高雑音環境下での自動音声認識では，このような音響雑音に影響されない情報を利用することによって認識率の向上が期待される．

これらはマルチモーダル音声認識と呼ばれる分野であるが，多くのマルチモーダル音声認識の研究では十分な認識精度の向上が実現できていない．この理由として，申請者は音声と画像の統合法に問題があると考える．従来のマルチモーダル音声認識における音声と画像の統合法は大きく２つに分けられる．１つ目は音声と画像でそれぞれ別々に認識を行い，それぞれで得られた認識結果を統合する結果統合である．２つ目は音声と画像を特徴量の段階で統合して認識に用いる初期統合である．ここで，ヒトの音声認識プロセスを考えた際，発話が雑音や発音の曖昧さなどの理由により聞こえづらかった部分でヒトは口唇の動きなどから補完して認識していると予想される．すなわち，観測した音声の情報に依存して，どのような音声と画像の情報を統合するのかを柔軟に変更していると考えられるのだが，従来のマルチモーダル音声認識の統合法ではこれが実装されていない．そこで，本研究では，観測音声信号に依存させて音声と画像の情報を適切に組み合わせる新しい音声と画像情報の統合法を提案し，特に高雑音環境下での認識率の向上を目指す．

本提案手法では，観測音声に依存させて適切に音声と画像の統合を実現するために，1) 雑音重畳音声特徴量の確率密度関数を，複数の正規分布の混合分布と仮定して予め学習し，2) 各正規分布ごとに，雑音重畳音声と画像からクリーン音声を推定する変換関数を用意，3) 最後に雑音重畳音声を観測したとき，それがどの正規分布から出力されたかの事後確率を重み付けとして，複数の変換関数を重ね合わせることで，クリーン音声を推定する枠組みを提案する．図 1，2 は観測音声の特徴量空間を簡易化して示している．従来手法では全ての観測特徴量に対して一律の変換を行うのに対し，提案手法では，観測特徴量によって異なる変換をかける．その結果，入力の音素や雑音の種類などに依存して適切な変換が選ばれることによって，適切にクリーン音声を推定することが可能となる．

本研究の特色は，観測音声に依存させて音声と画像の統合法を適切に変更している点と，画像情報を利用してクリーン音声を推定しそれを用いて認識を行っている点である．クリーン音声への変換を観測音声に依存させているため，ヒトに近い認識プロセスが実現され，精度向上に有効である．また，画像情報を用いてクリーン音声を推定する枠組みであるため，自動音声認識応用の他にも，例えば音声再合成を行うことにより，高雑音環境下での通話品質の向上などが実現できる．

本手法の有効性を示すために，日本語数字読み上げに対する認識実験による性能評価を行った．性能評価には，日本語数字読み上げの口唇画像と音声の対応が取れているコーパスである CENSREC-1-AV を利用した．その結果，従来手法の初期統合の一種である PCA を用いた手法を比較して，音声認識実験において約 25%のエラー削減率を得ることができ，非常に高い認識率を実現することができた．

目次

図目次

表目次

第1章

序論

1.1　本研究の背景

我々人間が他人とコミュニケーションを行う際，音声は非常に重要な位置を占めている．そのため，ユーザインターフェースとして音声を利用することは，自然な流れであろう．音声認識の研究の歴史は古く，20 世紀半ばから研究されてきた．1970 年代までは DP マッチング法による手法が主流であったが，隠れマルコフモデル (Hidden Markov Model ; HMM) による統計確率手法が使われるようになった．1990 年代になると国防省高等研究計画局（DARPA）などの国家プロジェクトによる大規模なコーパスの整備や，N-gram 法の提案により大語彙連続音声認識が可能となった．

近年，目覚ましいコンピュータの発展により，従来では難しかった大規模なデータを利用した統計的確率手法により，音声認識の精度は大きく向上している．未だ認識誤りは発生するものの，Google Chrome の音声認識エンジンや，iPhone に搭載されている Siri などのように商業利用されているケースも出て来ており，今後，ますますユーザインターフェースとしての利用が増えることが予想される．

しかし，実環境下では，音声認識は理想通りのパフォーマンスを発揮できることは限らない．これは，周辺の雑音などの影響により，機械が話者の音声を「聞き取りづらく」なるためである．話者とマイクが比較的近い場合は，この影響がさほど問題とならないことも多いため，現在の商業利用も多くは，マイクが口元にあるようなシステム，もしくは雑音のないクリーン環境での利用を想定した物が多い．しかし，ユーザインターフェースとしての更なる利用を考えた際，この問題を避けて通ることは出来ない．では，この「聞き取りづらさ」の問題をいかに克服すれば良いのだろうか．

この問題を解決するために，音声認識研究の分野では様々な研究がなされてきた．例えば，入力信号から直接雑音成分を除去する信号処理手法 [1]，特徴量ベースで雑音成分を除去する特徴量強調手法 [2, 3, 4]，音響モデルを雑音に対して適応するモデル適応手法 [5, 6, 7]，認識結果の仮説群をリランキングするリランキング手法 [8]，そして音声だけではなく画像などの副次的な情報を利用するマルチモーダル手法 [9] など多岐に渡る．これらの研究によって，クリーン音声に比較的近い環境では目覚ましい向上が得られたが，高雑音環境下では未だに低い認識率しか得られない．

さて，ここで，今一度人間の音声認識能力について焦点を当ててみよう．高雑音環境下では，人間も必ずしも確実に音声認識が可能であるわけではないのは，誰もが経験したことがあるであろう．しかし，一般に非常に騒音が大きな環境で，我々が他者とコミュニケーションを行う際は，相手の口唇の動きなどによって，ある程度相手が「いつ」喋っているか,「どのような音」を喋っているかを予測することが出来る．人間かは実空間で音声認識を行う際，耳から入る音だけではなく，これらの視覚情報などの副次的な情報を利用していると考えられる [10] ．これらの副次的な情報は特に，高雑音環境下での音声認識においてこそ有効に働くものであると考えられる．

1.2　本研究の目的

本稿では口唇動画像を用いたマルチモーダル手法に焦点を当てて，高雑音環境下における音声認識の精度を向上を計る．マルチモーダル手法は，音声と同時に様々な情報を利用することによって認識精度を向上させる．これは顔の表情や口唇の動きなどと発話には相関があると考え，それらを認識の際に利用するものである．これは口唇動画像に限らず，筋電等を利用する研究も存在する [11]．また，音声認識に直接利用するだけでなく音声区間検出に用いる例などもある [12]．マルチモーダル音声認識は大きく分けて音声と画像別々に認識して得られた尤度を統合する結果統合法 [28, 29] と特徴量ベースで結合して認識を行う初期統合法 [9, 26, 27] の 2 種類に大別することができる．

結果統合は音声のモデルと画像のモデルそれぞれから得られる尤度を重みづけにより統合する．結果統合では，デコーディングを複数回行ない，それらを重みつけする必要があるのに対し，初期統合は特徴量を変更するだけで，モデルの学習が比較的簡単であるという性質がある．この性質により，モデルの適応などと組み合わせやすい利点がうまれるため，本稿ではこの初期統合に注目する．初期統合の問題点は特徴量をどのような基準で統合するかであり，この基準は，雑音の種類や大きさ，発話の音素の種類などによって適切に変更するべきであると考えられる．特に高雑音環境下では，判別しづらい音素などに傾向があるはずである．

そこで，本研究では，これを実現するために，区分的線形変換を用いた新しい初期統合法を提案する．これは，劣化音声特徴量をガウス混合分布 (Gaussian Mixture Model : GMM) を用いることで多くのクラスに分類し，それぞれのクラスに所属する音声特徴量と画像特徴量の連結ベクトルに対する最適な線形変換を推定する．これによって, クラスに依存した最適な基準で音声と画像の特徴量を利用することができる．本手法は劣化音声特徴量からクリーン音声特徴量を区分的線形変換によって推定する特徴量強調手法である SPLICE を参考としている．

1.3　本論文の構成

本論文は，全 7 章で構成される．第 **2** 章では，特に本研究で用いる音声情報処理の基礎と従来の音声認識システムについて説明を行う．第 **3** 章では，雑音が音声認識システムに与える影響について論じ，従来の耐雑音処理技術について紹介を行う．第 **4** 章では，本研究の核となる区分的線形変換について，SPLICE をベースにその性質について説明を行う．第 **5** 章では提案手法について説明を行い，第 **6** 章で本提案手法の有効性を示すための実験について述べる．最後に第 **7** 章で本論文をまとめ，今後の展望について述べる．

第2章

音声情報処理の基礎と音声認識技術

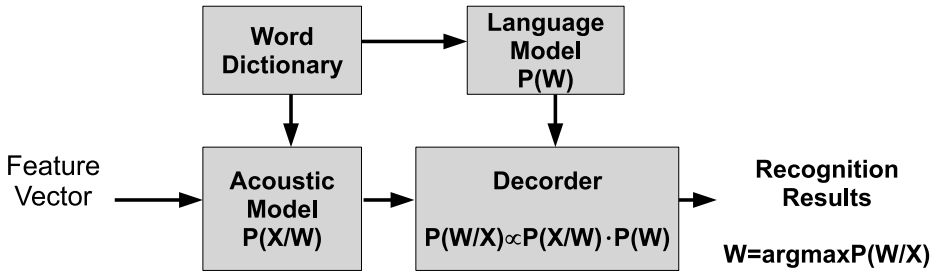

図 2.1: 音声認識システム

2.1　はじめに

本章では，音声認識システムの全体像について紹介する．音声認識は一般に，入力音声から信号処理・音声分析により特徴量系列 $\boldsymbol{X}$ を求める (図 (2.1))．その後，特徴量系列 $\boldsymbol{X}$ に対して事後確率 $P(\boldsymbol{W}|\boldsymbol{X})$ を最大とする単語系列 $\boldsymbol{W}$ を見つける問題として定式化される．まず音声認識において用いられる音響特徴量であるケプストラム係数について述べる．その後，音響モデルと言語モデルについて概説を行う．

2.2　音響特徴量

音声の生成メカニズムは声帯の振動や乱流などによって生じる音源と，声道による伝達特性を持つ調音フィルタによるソースフィルタモデルとして考えることができる．音声の最小単位である音素は，音源の種類（破裂音，声帯振動，乱流等）とフィルタの形状で決定されるが，特に音源の性質は，声の高さや大きさ等の音声の韻律的な性質に影響している．そのため，音源の情報を除いた，声道フィルタの伝達特性に相当する情報を抽出することで，音声認識に有効な音響特徴量が得られると考えられる．この考えに基づいて設計された特徴量がケプストラムである．本節では，まずケプストラム分析について述べ，その後，音声認識において用いられることの多い特徴量であるメル周波数ケプストラム係数について紹介を行う．

2.2.1　ケプストラム

音声は，声帯の振動や摩擦による乱流等の音源信号が調音フィルタの伝達特性によって音韻情報が付与された物であり，音素の音響的な特徴は，主に調音フィルタの振幅伝達特性によって担われている．そのため，音声認識に有効な特徴量を得るには，音声から音源成分を除去し，調音フィルタの伝達特性に相当する情報を抽出すれば良いことがわかる．ケプストラム（Cepstrum）は，声帯の振動成分と調音フィルタの伝達特性を比較した際に，調音フィルタの伝達特性が対数振幅スペクトルの包絡に相当することを利用して抽出する．

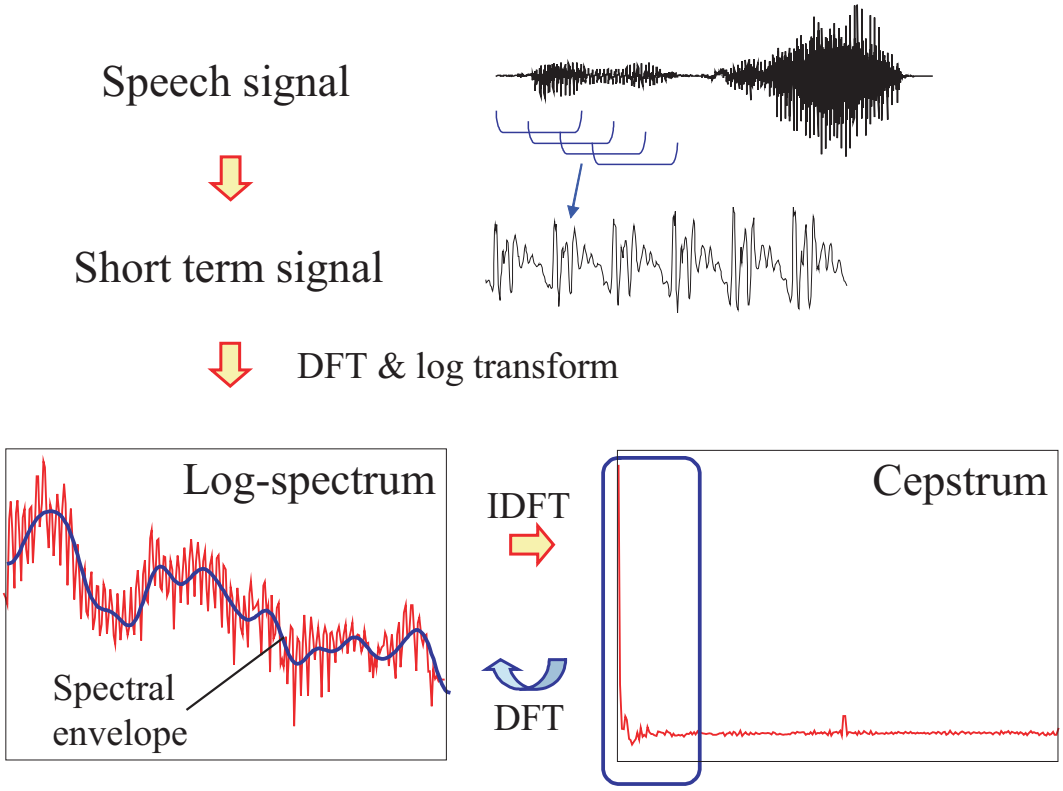

図 2.2: ケプストラム

図 (2.2) にケプストラム分析の概形を示す．音声は窓関数との畳み込みを行った後，短時間フーリエ変換により，スペクトルを得ることができる．音源信号のスペクトラムを $G(e^{j\omega})$，調音フィルタの伝達特性を $H(e^{j\omega})$ とすると，音声信号のスペクトル $S(e^{j\omega})$ は，

$$S(e^{j\omega}) = G(e^{j\omega}) \cdot H(e^{j\omega}) \tag{2.1}$$

と表すことができる．ここで，(対数) 振幅スペクトルを考えると，

$$|S(e^{j\omega})| = |G(e^{j\omega})| \cdot |H(e^{j\omega})| \tag{2.2}$$

$$\log|S(e^{j\omega})| = \log|G(e^{j\omega})| + \log|H(e^{j\omega})| \tag{2.3}$$

となる．したがって，音声信号の対数振幅スペクトルは，音源と調音フィルタの対数振幅スペクトルの和で表すことができることがわかる．ここで，対数振幅スペクトルの微細構造は音源成分を，包絡は調音フィルタの成分が表れていることが知られており，これを時間信号と考えて，さらに逆フーリエ変換を行うことで，低い周波数帯域に調音フィルタに相当する情報が集中すると考えることができる．つまり，対数振幅スペクトルの逆フーリエ変換がケプストラム係数であり，ケプストラム係数の低次項からは韻律に相当する情報が取り除かれていると考えることができる．

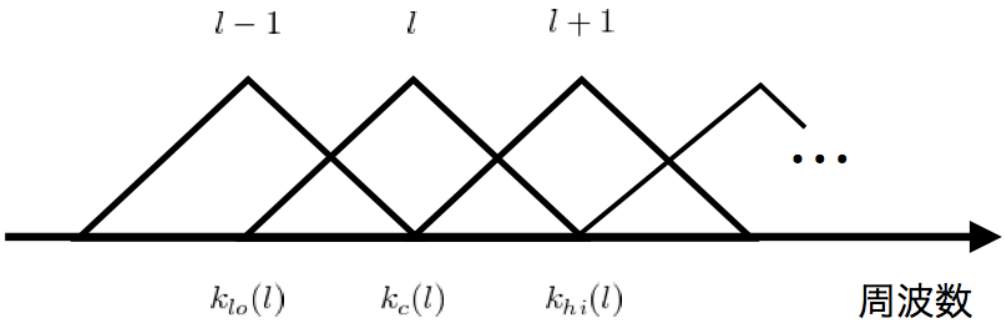

図 2.3: メル周波数フィルタバンク

2.2.2　メル周波数ケプストラム係数

一般に，音声認識では人間の聴覚特性にあわせて調整したメル周波数領域でケプストラム分析を行うことによって得られるメル周波数ケプストラム係数（Mel-Frequency Cepstrum Coefficient ; MFCC）を用いることが多い．メル周波数は

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.4}$$

により研鑽することができる．MFCC は，図 (2.3) のように周波数軸上に L 個の三角窓を配置し，フィルタバンク分析により得ることが出来る．音声スペクトルを $S'(k)$ とするとし，各窓の中心点 $k_c(l)$ をメル周波数軸上で等間隔に配置すると，L 個の窓から得られる出力は

$$m(l) = \sum_{k=lo}^{hi} W(k;l)|S'(k)| \quad (l = 1, \cdots, L) \tag{2.5}$$

$$W(k;l) = \begin{cases} \frac{k-k_{lo}(l)}{k_c(l)-k_{lo}(l)} & \{k_{lo} \le k \le k_c(l)\} \\ \frac{k_{hi}(l)-k}{k_{hi}(l)-k_c(l)} & \{k_c \le k \le k_{hi}(l)\} \end{cases} \tag{2.6}$$

となる．その後，得られた L 個のパワーを離散コサイン変換（DCT）することで，MFCC が求まる．

$$c_{mfcc}(i) = \sqrt{\frac{2}{N}} \sum_{l=1}^{L} \log m(l) \cos\left\{\left(l - \frac{1}{2}\right)\frac{j\pi}{L}\right\} \tag{2.7}$$

2.2.3　Δケプストラム

音響的な特徴はフレーム単位の値だけではなく，その動きに表れると考えられる．そこで，特徴量として MFCC の一次微分，ないしは二次微分の値を導入することが一般的に行われる．これは一般的にΔケプストラムと呼ばれ，音声認識では比較的良く用いられる．

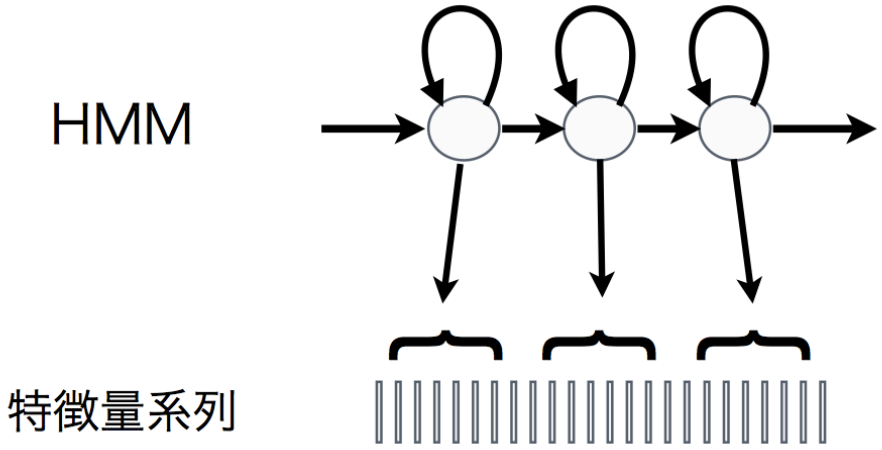

図 2.4: 隠れマルコフモデル

2.3　音響モデル

音響モデルは観測特徴量系列がどの音素（もしくは単語）であるかを計算するためのものである．音声は系列データであるため，系列でモデル化しなければならない．本節では，時系列データのモデル化のため，音声認識において用いられる left-to-right 型の隠れマルコフモデル（Hidden Markov Model ; HMM）について説明を行う．その後，連結学習について述べ，モデルの構築スケールについて述べる．

2.3.1　隠れマルコフモデル

HMM は系列データためのモデルであり，複数の状態と状態間の遷移確率，そして各状態からの出力の分布を持つ．図 (2.4) に HMM の概形を示す．音声認識では，一般的に left-to-right 型の HMM が用いられる．HMM は状態遷移を隠れ変数に持つモデルであり，学習は Baum-Welch アルゴリズムによって行う．これは EM アルゴリズムの HMM への拡張である．特徴量系列を $\boldsymbol{X}$，モデルパラメータを Θ とし，

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \sum_{l} \log P(\boldsymbol{X}_l|\Theta) \tag{2.8}$$

のように最尤基準でモデルパラメータを学習する．ここで，l はデータインデクスである．隠れ状態系列を $\boldsymbol{s}$ とすると，

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \sum_{l} \log \sum_{\boldsymbol{s}} P(\boldsymbol{X}_l, \boldsymbol{s}|\Theta) \tag{2.9}$$

$$= \underset{\Theta}{\operatorname{argmax}} \sum_{l} \log \sum_{\boldsymbol{s}} P(\boldsymbol{X}_l|\boldsymbol{s}, \Theta) P(\boldsymbol{s}|\Theta) \tag{2.10}$$

これを直接計算することは困難であるため，下限を保証し反復計算を行うことで最大化を行う．Q 関数は

$$Q(\Theta|\Theta^{(n)}) = \sum_{l}\sum_{\boldsymbol{s}} P(\boldsymbol{s}|\boldsymbol{X}_l, \Theta^{(n)}) \log P(\boldsymbol{X}_l|\boldsymbol{s}, \Theta) P(\boldsymbol{s}|\Theta) \tag{2.11}$$

となる．

さて，実際には状態系列は非常に膨大であるため，$P(\boldsymbol{s}|\boldsymbol{X}_l, \Theta^{(n)})$ を全ての状態系列について計算するのは困難である．そこで，Baum-Welch アルゴリズムでは，効率的に計算するために前向き・後ろ向きアルゴリズムを一般的に用いる．データ l を前向き確率 $\alpha_l(t,i)$ は，時刻 t において状態 i に至る確率であり，

$$\alpha_l(t,i) = P(s_t = i, \boldsymbol{x}_{l,1}, \boldsymbol{x}_{l,2}, \ldots, \boldsymbol{x}_{l,t}|\Theta) \tag{2.12}$$

となる．また，時刻 t において，状態 j を出発して時刻 $T+1$ に終状態 M に辿り着く確率である後ろ向き確率 $\beta_l(t,j)$ は，

$$\beta_l(t,j) = P(\boldsymbol{x}_{l,t+1}, \boldsymbol{x}_{l,t+2}, \ldots, \boldsymbol{x}_{l,T}|s_t = j, \Theta) \tag{2.13}$$

となる．これらを用いると，時刻 t において状態 m を通過する系列の出現確率は

$$P(s_t = m|\boldsymbol{X}_l, \Theta) = \frac{\alpha_l(t,m)\beta_l(t,m)}{\alpha_l(T+1,M)} = \psi_l(t,m) \tag{2.14}$$

となる．

各状態における特徴量分布を正規分布と仮定すると，求めるべきパラメータは状態遷移確率 $a(m)$ と各状態における特徴量分布の平均 $\boldsymbol{\mu}_m$ と分散 $\boldsymbol{\sigma}_m^2$ である．最終的に，$b_m(\boldsymbol{x})$ を状態 m の特徴量分布から特徴量 $\boldsymbol{x}$ が出力される確率とすると，それぞれの更新式は

$$\hat{a}(m) = \frac{\sum_l \left[\frac{1}{\alpha_l(T+1,M)} \sum_t \alpha_l(t,m) a(m) b_m(\boldsymbol{x}_{lt}) \beta_l(t,m)\right]}{\sum_l \sum_t \psi_l(t,m)} \tag{2.15}$$

$$\hat{\boldsymbol{\mu}}_m = \frac{\sum_l \sum_t \psi_l(t,m) \boldsymbol{x}_{lt}}{\sum_l \sum_t \psi_l(t,m)} \tag{2.16}$$

$$\hat{\boldsymbol{\sigma}}_m^2 = \frac{\sum_l \sum_t \psi_l(t,m) (\boldsymbol{x}_{lt} - \hat{\boldsymbol{\mu}}_m)(\boldsymbol{x}_{lt} - \hat{\boldsymbol{\mu}}_m)^\top}{\sum_l \sum_t \psi_l(t,m)} \tag{2.17}$$

となる．音声認識の場合，一般的には特徴量分布を混合正規分布（Gaussian Mixture Model ; GMM）や多層パーセプトロンによる識別モデルによって表現することが多い．

また，構築する HMM の種類によってワード型とサブワード型 HMM に分類することができる．サブワード型は音素などの音声の最小単位で HMM を学習するのに対して，ワード型は単語単位で HMM を構築する．これらはタスクによって向き不向きが変わるが，サブワード型は少量の HMM で多様な表現ができるため，大語彙音声認識は音素単位での HMM が使われる．

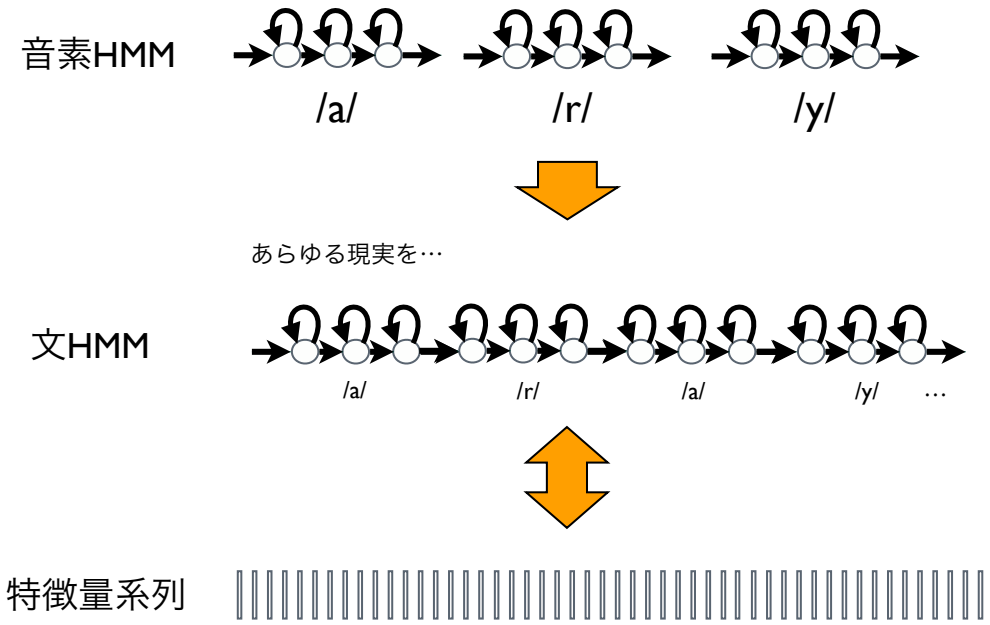

図 2.5: 連結学習

2.3.2　連結学習

実際の学習データは，文章を読み上げた音声とそれに対応するテキストが与えられることが多い．しかし，音素 HMM を構築する場合，音素ラベルとそれに対応する音声が本来必要である．このミスマッチを解消するため，音素 HMM を連結し文 HMM を作成し，それを用いて HMM の学習を行う．図 (2.5) のように音素 HMM のパラメータを連結した文 HMM と特徴量系列を用いてモデルパラメータを学習することにより，音素単位での学習データがなくとも学習することができる．

2.3.3　音素文脈

音素はそれぞれで独立ではなく，前後の音素の影響を大きく受けることが知られている．これは，調音結合と呼ばれ，音声の認識を困難にしている要因の 1 つである．そのため，連結学習で単純に音素を当てはめるのでは，前後の音素の違うものも全て同一の音素であるとして学習が行われてしまう．これに対処するため，前後の音素を考慮した 3 つ組み音素（triphone）単位として HMM のモデル化を行う．日本語の音素はおおよそ 40 種類と言われ，40 種類の音素に対応するモデルを学習すれば任意の単語の認識が可能となる．．しかし，前後の音によって音素の音が影響をうけてしまう調音結合が起こるため，一般的には，直近の前後の音素を考慮する triphone を用いられることが多い．また，さらに細かなモデ

ル化のため，4 つ組み以上も考慮する場合もある．逆に，学習データが少量の場合は，スパース性の問題から monophone で行われる．

しかし，triphone は膨大なクラス数となるため，このままではスパースなため，学習データ中に一度も出ない 3 つ組み音素もある．そこで，HMM を決定木などによりクラスタリングすることにより対処することが一般に行われる．

2.4　言語モデル

認識対象の単語数や文が限定的である場合は，文法は連結学習と同様に音素 HMM や単語 HMM を連結した文 HMM を構築することによって表現することができる．しかし，大語彙連続音声認識では，候補の文が膨大であるため，このようなアプローチには限界がある．そこで，大規模コーパスを用いて自動的にモデルを作成する統計的言語モデルが用いられる．本節では，代表的な統計的言語モデルである N-gram モデルについて紹介する．

2.4.1　N-gram

N-gram モデルでは単語列 $w_1, w_2, \ldots, w_n$ が与えられた時の，その出現確率 $P(w_1, w_2, \ldots, w_n)$ を

$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_{i-N+1}, \ldots, w_{i-1}) \tag{2.18}$$

と近似を行う．これは，つまり，i 番目の単語 w_i の生成確率が，直前の $N-1$ 単語 $w_{i-N+1}, \ldots, w_{i-1}$ にのみ依存すると仮定する．ここで，$N = 1$ の場合がユニグラム，$N = 2$ の場合がバイグラム，$N = 3$ の場合がトライグラムと呼ばれる．N-gram の学習は最尤推定によって行う．そのため，学習データに存在しないものについては確率が 0 となる．言語モデルは非常にスパースであるため，確率が 0 となるものが発生しやすく，これが認識誤りを引き起こす．そこで，確率の高いものを下げ，確率の小さなものを上げる smoothing を行うことによって，一般にこの問題に対処する．

2.5　まとめ

本章では音声認識システムの全体像について述べた．音声認識は特徴量抽出，音響モデル，言語モデルに分けて考えることができる．次章ではこの音声認識システムを拡張した口唇動画像を用いたマルチモーダル音声認識システムについて述べる．

第3章

マルチモーダル音声認識

3.1　はじめに

前章では音声認識システムの概要について述べた．本章では，マルチモーダル音声認識システムについて述べる．音響特徴量は雑音の影響で大きく歪むことが知られており，これが実環境下での音声認識において認識率の低下を引き起こす原因となっている．これは，本質的には，モデルと雑音で歪んだ音声特徴量分布との間のミスマッチが問題となる．マルチモーダル音声認識は，音響雑音の影響を受けない画像などの情報を併用することにより，相対的にミスマッチを抑制することができる．

本章では，口唇動画像情報を利用したマルチモーダル音声認識の手法について説明する．マルチモーダル音声認識において，どのような画像特徴量が発話を適切に表現しているか，そして画像特徴量をどのように認識プロセスに組み込むかが問題となる．

3.2　画像特徴量

画像情報を利用したマルチモーダル音声認識において用いられる画像特徴量は Appearance ベース，Shape ベース，そしてその両方を用いたものの大きく 3 種類に分かれる．本節では，Appearance ベースと Shape ベースの特徴量について紹介した後，それらを組み合わせた代表的な特徴量抽出アルゴリズムとして Active Appearance Model について紹介する．

3.2.1　Appearance ベース特徴

Appearance ベースは画像のすべてのピクセルが発話と相関があると考え，対象領域内のピクセル値を用いる．最も一般的なのは，主成分分析によって領域内のピクセル値を表現することによって特徴量の次元を削減しつつ効率的に扱うことができる [13, 14, 15]．また，画像を周波数情報と考え，離散コサイン変換を行うことにより特徴量を抽出する手法もある [16, 17]．

3.2.2　Shape ベース特徴

Shape ベースは口唇やその他の顔のパーツの形が発声に相関を持っていると考え，例えば口唇の高さや幅などから特徴量を抽出する．単純に口唇の形状をそのまま特徴量として用いる場合 [18, 19, 20] だけでなく，統計的な口唇の形状をモデル化してパラメータにより表現することも行われている [21, 22, 23, 24]．これは比較的照明の変化などに強いという特徴があるが, 口唇の輪郭を正確に抽出する必要 があるために, 計算量が高くなるなどの問題がある．

3.2.3　Active Appearance Model

Active Appearance Model[25] は，Appearance ベースと Shape ベースを組み合わせた特徴量であり，高速に計算が可能である．特徴点の形状 $\boldsymbol{s}$ と輝度値 $\boldsymbol{g}$ を主成分分析により次元

圧縮を行い，低次元のパラメータによりモデルを表現する手法である．特徴点には Haar-like 特徴[1]を用いる．形状 $\boldsymbol{s}$ と輝度値 $\boldsymbol{g}$ はそれぞれ，

$$\boldsymbol{s} = (x_1, y_1, \ldots, x_n, y_n)^\top \tag{3.1}$$
$$\boldsymbol{g} = (g_1, \ldots, g_m)^\top \tag{3.2}$$

と表すことが出来る．ここで，x,y はそれぞれの特徴点の座標である．ただし，g は平均形状 $\overline{\boldsymbol{s}}$ に画像を正規化したときの $\overline{\boldsymbol{s}}$ 内部での各画素の輝度値である．これらに対して，それぞれ主成分分析を行う．

$$\boldsymbol{s} = \overline{\boldsymbol{s}} + \boldsymbol{P}_{\boldsymbol{s}}\boldsymbol{b}_{\boldsymbol{s}} \tag{3.3}$$
$$\boldsymbol{g} = \overline{\boldsymbol{g}} + \boldsymbol{P}_{\boldsymbol{g}}\boldsymbol{b}_{\boldsymbol{g}} \tag{3.4}$$

$\overline{\boldsymbol{g}}$ は平均輝度値，$\boldsymbol{P}_{\boldsymbol{s}}$，$\boldsymbol{P}_{\boldsymbol{g}}$ は $\overline{\boldsymbol{s}}$，$\overline{\boldsymbol{g}}$ からの偏差を主成分分析して得られる固有ベクトルである．また，$\boldsymbol{b}_{\boldsymbol{s}}$，$\boldsymbol{b}_{\boldsymbol{g}}$ はそれぞれ shape パラメータ，texture パラメータと呼び，平均からの変化を表すパラメータである．その後，形状と輝度値にも相関があると考え，$\boldsymbol{b}_{\boldsymbol{s}}$ と $\boldsymbol{b}_{\boldsymbol{g}}$ をさらに主成分分析する．

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{W}_{\boldsymbol{s}}\boldsymbol{b}_{\boldsymbol{s}} \\ \boldsymbol{b}_{\boldsymbol{g}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}_{\boldsymbol{s}}\boldsymbol{P}_{\boldsymbol{s}}^\top(\boldsymbol{s} - \overline{\boldsymbol{s}}) \\ \boldsymbol{P}_{\boldsymbol{g}}^\top(\boldsymbol{g} - \overline{\boldsymbol{g}}) \end{pmatrix} = \boldsymbol{Q}\boldsymbol{c} \tag{3.5}$$
$$\boldsymbol{g} = \overline{\boldsymbol{g}} + \boldsymbol{P}_{\boldsymbol{g}}\boldsymbol{b}_{\boldsymbol{g}} \tag{3.6}$$

最終的に，$\boldsymbol{c}$ を用いて $\boldsymbol{s}$,$\boldsymbol{g}$ を表現すると

$$\boldsymbol{s}(\boldsymbol{c}) = \overline{\boldsymbol{s}} + \boldsymbol{P}_{\boldsymbol{s}}\boldsymbol{W}_{\boldsymbol{s}}^{-1}\boldsymbol{Q}_{\boldsymbol{s}}\boldsymbol{c} \tag{3.7}$$
$$\boldsymbol{g}(\boldsymbol{c}) = \overline{\boldsymbol{g}} + \boldsymbol{P}_{\boldsymbol{g}}\boldsymbol{Q}_{\boldsymbol{g}}\boldsymbol{c} \tag{3.8}$$

として表現することが可能となる．ただし，$\boldsymbol{W}_{\boldsymbol{s}}$ は形状ベクトルと輝度値ベクトルの単位の違いを正規化する行列，$\boldsymbol{Q}$ は固有ベクトル，$\boldsymbol{c}$ は形状と輝度値の両方を制御するパラメータで combined パラメータと呼ぶ．このようにモデルを構築することで，1 つのパラメータ $\boldsymbol{c}$ を操作することで様々な状態を表現することが可能となる．図は AAM を用いて顔全体のパーツのパラメータ抽出を行ったものである．

3.3　統合法

マルチモーダル音声認識において，最も重要となるのは，音声と画像のストリームをいかに統合するかである．一般に，統合法は初期統合法 [9, 26, 27] と結果統合法 [28, 29] の 2 つに大別することができる．本節では，それぞれの統合について説明を行う．

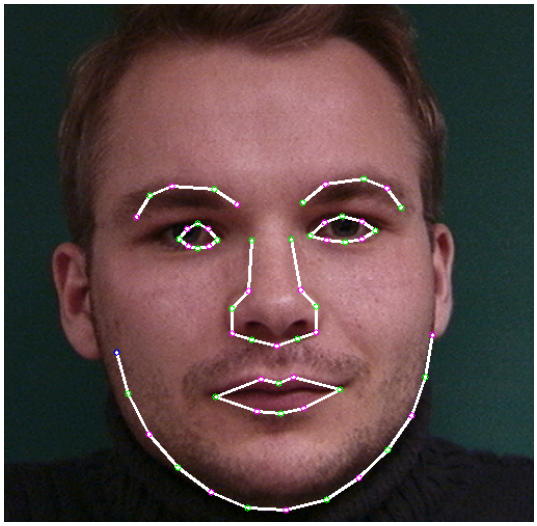

図 3.1: Active Appearance Model を用いた顔パーツ抽出

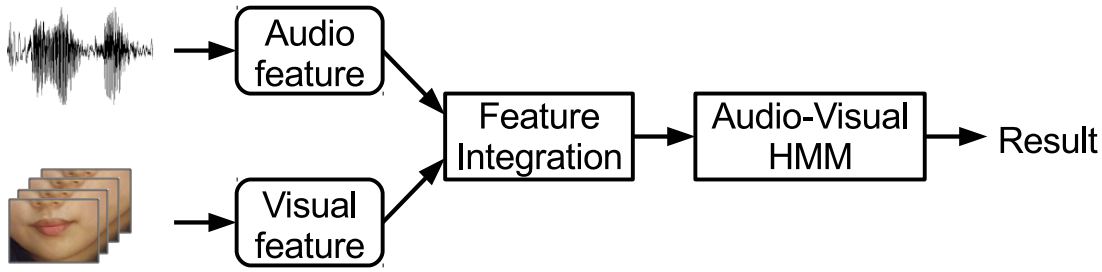

図 3.2: 初期統合のフロー図

3.3.1　初期統合法

初期統合法は，入力特徴量の段階で音声と画像のストリームを統合を行う．図 (3.2) に初期等合法のフロー図を示す．y を音声特徴量，i を画像特徴量とすると，連結特徴量 m は

$$\boldsymbol{m} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{i} \end{bmatrix} \tag{3.9}$$

となる．初期統合法では，この連結特徴量を用いて HMM モデルを構築する．しかし，連結特徴量は次元数が大きいため，更に 主成分分析（Principle Component Analysis ; PCA）や線形判別分析（linear discriminant analysis ; LDA）を式 (2) のようにかけて次元圧縮したベクトル n を利用することが多い．

$$\boldsymbol{n} = \boldsymbol{Q}\boldsymbol{m} \tag{3.10}$$

[1]近接する矩形領域の明度差を用いることで得られる．照明条件などのノイズに強い．

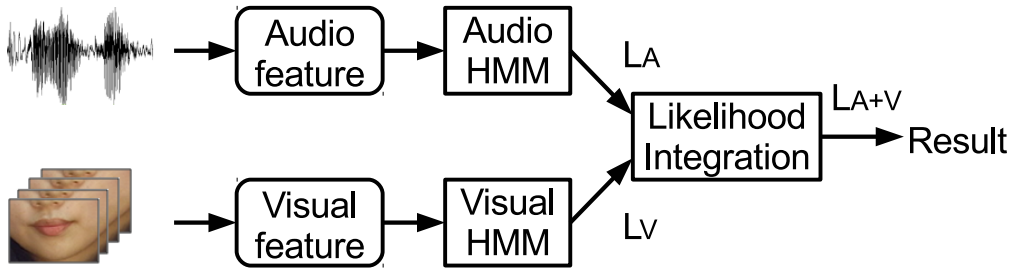

図 3.3: 結果統合のフロー図

ここで $\boldsymbol{Q}$ は任意の次元圧縮手法により学習した変換行列である．初期統合は，特徴量を統一的に扱える反面，それぞれの特徴量に起因するオーダーの違いから正規化の問題などが発生する．

3.3.2　結果統合法

結果統合は音声と画像の各ストリームで認識を行った後，それぞれの認識結果を Adaboost などによって統合する．図 (3.3) に初期等合法のフロー図を示す．結果統合は複数の認識プロセスを介する必要があるため，計算量が大きくなるという問題がある．また，単純に認識結果のみを統合する場合では，各フレームで見ると，音声と画像で異なる状態割り当て（アライメント）が発生する．これを改善するため，マルチストリーム HMM という状態単位での尤度統合法が行われている [30]．

i)　マルチストリーム HMM

マルチストリーム HMM は，音声と画像のアライメントを一致させるため，各ストリームの HMM の状態を同期する．図 (3.4) にマルチストリーム HMM の概形を示す．ある時間 t における観測特徴量 $\boldsymbol{x}_t$ に対する対数尤度は各ストリームから得られる対数尤度の線形和で表現される．音声ストリームから得られた対数尤度を $L_{a,t}$，画像ストリームから得られた対数尤度を $L_{v,t}$ とすると，観測特徴量 $\boldsymbol{x}_t$ に対する対数尤度は

$$L_{av,t} = \lambda_a L_{a,t} + \lambda_v L_{v,t} \quad \text{where} \quad \lambda_a + \lambda_v = 1 \quad \lambda_a, \lambda_v \geq 0 \tag{3.11}$$

となる．マルチストリーム HMM の最大の特徴は，音声と画像のストリームにおけるアライメントが一致する点である．一般にマルチストリーム HMM の学習は，音声 HMM をあらかじめ学習しておき，それを用いて音声のアライメントを行った後，画像ストリームをアライメント情報を用いて学習する．

コーパスとして CENSREC-1-AV を用いた単語認識結果を表 (3.1) に示す．音声対雑音比 (SNR) が小さな雑音が大きい環境では，音声のみによる単語認識率が 57.34%であるのに対し，マルチストリーム HMM は 72.35%を示す．ここで，マルチストリーム HMM における認識率の隣の添字は，最も認識率の良い場合の音声ストリーム統合重み λ_a である．雑音環境下では，0.7 と画像ストリームの割合が増加していることがわかる．

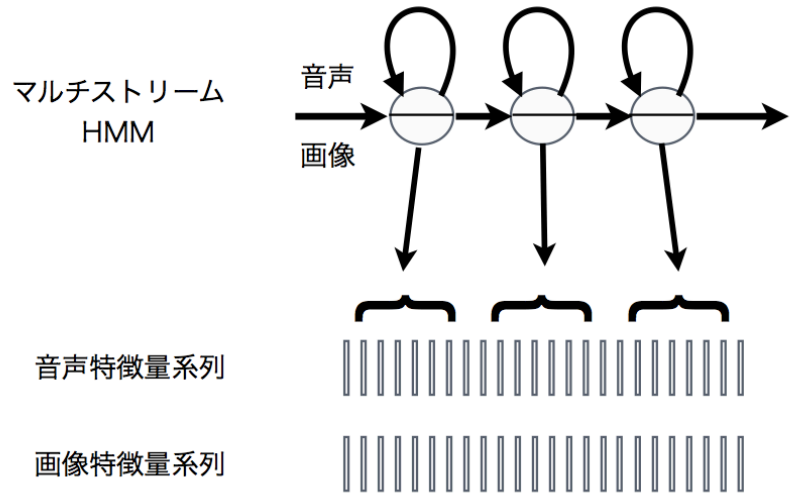

図 3.4: マルチストリーム HMM

表 3.1: マルチストリーム HMM を用いた単語認識結果 %（CENSREC-1-AV）

	clean data	SNR=-5dB (expressway noise)
visual only	32.73	
audio only	99.61	57.34
audio and visual	99.36 [0.9]	72.35 [0.7]

3.4　まとめ

本章では，マルチモーダル音声認識について述べた．まず，用いられる画像特徴量について触れた．その後，画像情報と音声情報の統合法について述べ，マルチストリーム HMM について紹介した．次章では，本提案手法の核となる区分的線形変換について，区分的線形変換を用いた代表的な特徴量強調手法である SPLICE をベースにその性質について述べる．

第4章

区分的線形変換とその拡張性

4.1　はじめに

前章では，雑音の影響によって音声認識の精度が著しく低下することに対処するためのアプローチの 1 つであるマルチモーダル音声認識について紹介した．先に述べた通り，画像情報を音声認識に利用することを考えた際，重要になるのは音声と画像のドメインの違いとそれに起因する統合法である．そのため，特徴量強調に画像情報を利用しようと考えた際，ドメインの違いを上手く吸収する枠組みが必要である．そこで，入力特徴量に応じて柔軟な変換が可能である区分的線形変換が有効であると考えられる．

本章ではまず，区分的線形変換を利用した手法の代表例である SPLICE[4] を用いてその特性を説明する．これを踏まえた上で，拡張性について事例を紹介して紹介する．

4.2　区分的線形変換を用いた特徴量強調

まず，区分的線形変換を用いた特徴量強調手法である SPLICE（Piecewise LInear Compensation for Environments）の説明を行い，その特性について考察を行う．雑音の影響により，観測音声のメル周波数ケプストラム係数（MFCC）がクリーン環境における音声の MFCC と比べ大きく歪み，モデルとのミスマッチを引き起こす原因となることが知られている．これを解決するために，特徴量強調は，観測音声特徴量 $\boldsymbol{y}$ が得られた時に，クリーン音声特徴量 $\boldsymbol{x}$ を推定する．モデルのパラメータを $\boldsymbol{\Theta}$ とすると事後確率 $P(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\Theta})$ を最大化することを考える．

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, P(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\Theta}) \tag{4.1}$$

ここで $\hat{\boldsymbol{x}}$ は推定されたクリーン音声特徴量である．雑音の種類にも依存するが，基本的に劣化音声特徴量とクリーン音声特徴量との関係は非線形で表されるため，非常に複雑である．そのため，VTS-based のようにその関係性を得るためには雑音の情報が必要となるが，雑音の推定精度の影響を大きく受けてしまうという問題点がある．対して，SPLICE は，「近い特徴量は同じ線形変換でクリーン音声特徴量に変換可能である」という仮定をおくことで雑音の情報を介さずにクリーン音声特徴量を推定する試みである．図 (4.1) のように，単なる線形変換では，全ての特徴量空間に対して同一な変換を行うのに対し，SPLICE では特徴量空間でクラスタリングを行い，各クラスにおいてそれぞれ異なる変換を行うことで複雑な表現が可能となる．

4.2.1　SPLICE の定式化

まず, 特徴量空間での近さを基準に複数のクラスに分類する．クラスインデックスを k とすると

$$P(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\Theta},k) \;=\; \mathcal{N}(\boldsymbol{x};\boldsymbol{A}_k\begin{bmatrix}1\\ \boldsymbol{y}\end{bmatrix},\boldsymbol{\Sigma}_k) \tag{4.2}$$

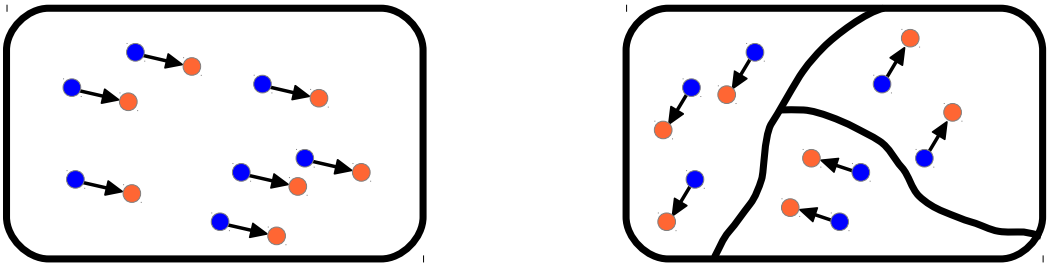

図 4.1: 特徴量空間のクラスタリングと変換

として平均が観測特徴量に対する線形変換を持つガウス分布で近似する．k は観測することができないため，周辺分布を最大化することを考える．

$$\begin{aligned}\hat{\boldsymbol{x}} &= \underset{\boldsymbol{x}}{\operatorname{argmax}} \sum_k P(\boldsymbol{x}, k|\boldsymbol{y}, \boldsymbol{\Theta}) \\ &= \underset{\boldsymbol{x}}{\operatorname{argmax}} \sum_k P(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Theta}, k) P(\boldsymbol{y}|\boldsymbol{\Theta}, k) P(k|\boldsymbol{\Theta}) \end{aligned} \tag{4.3}$$

ここで，$P(\boldsymbol{y}|\boldsymbol{\Theta}, k)$ をガウス分布，$P(k|\boldsymbol{\Theta})$ をコンポーネントごとの重みと考えることで，

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmax}} \sum_k P(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Theta}, k) P(\boldsymbol{y}|\boldsymbol{\Theta}, k) P(k|\boldsymbol{\Theta}) \tag{4.4}$$

$$P(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Theta}, k) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{A}_k \boldsymbol{y}, \boldsymbol{\Sigma}_k) \tag{4.5}$$

$$P(\boldsymbol{y}|\boldsymbol{\Theta}, k) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) \tag{4.6}$$

$$P(k|\boldsymbol{\Theta}) = \pi_k \tag{4.7}$$

としてモデル化を行う．このモデルは EM-algorithem を用いて学習が可能である．Q 関数は E を期待値演算子とすると，n を EM-algorithem のイテレーション数として

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)}) = \mathrm{E}_{k|\boldsymbol{x},\boldsymbol{y},\boldsymbol{\Theta}^{(n)}}[\log P(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Theta}, k) P(\boldsymbol{y}|\boldsymbol{\Theta}, k) P(k|\boldsymbol{\Theta})] \tag{4.8}$$

とおくことが出来る．学習データのインデックスを l，フレームを t とすると，
E-step:

$$P(k|\boldsymbol{x}_{lt}, \boldsymbol{y}_{lt}, \boldsymbol{\Theta}^{(n)}) = \frac{P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}^{(n)}, k) P(\boldsymbol{y}_{lt}|\boldsymbol{\Theta}^{(n)}, k) P(k|\boldsymbol{\Theta}^{(n)})}{\sum_{k'} P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}^{(n)}, k') P(\boldsymbol{y}_{lt}|\boldsymbol{\Theta}^{(n)}, k') P(k'|\boldsymbol{\Theta}^{(n)})} \tag{4.9}$$

M-step:

$$\begin{aligned}\boldsymbol{\Theta}^{(n+1)} &= \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \sum_l \sum_t Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)}) \\ &= \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \sum_l \sum_t \sum_k P(k|\boldsymbol{x}_{lt}, \boldsymbol{y}_{lt}, \boldsymbol{\Theta}^{(n)}) \log P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}, k) P(\boldsymbol{y}_{lt}|\boldsymbol{\Theta}, k) P(k|\boldsymbol{\Theta}) \end{aligned} \tag{4.10}$$

の繰り返しにより求めることができる．

認識時は，学習されたパラメータ群を用いて

$$\hat{\boldsymbol{x}}_{lt} = \sum_k P(k|\boldsymbol{x}_{lt}, \boldsymbol{y}_{lt}) \boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y}_{lt} \end{bmatrix} \tag{4.11}$$

としてクリーン音声特徴量を推定する．ただし，$P(k|\boldsymbol{x}_{lt}, \boldsymbol{y}_{lt}, \boldsymbol{i}_{lt}, \boldsymbol{\Theta})$ はクリーン音声特徴量に依存するため，これの計算は EM-algorithm などを用いることによってしか計算することは出来ない．しかし，EM-algorithm は大きな計算量を持ち，収束まで時間がかかる場合も十分考えられる．そこで，SPLICE では特徴量空間での局所領域の設定に対する近似を導入することでこれを回避する．

4.2.2　特徴量空間における局所領域の設定

SPLICE では，EM-algorithm の E-step での計算に相当する，データの各コンポーネントに対する重み $P(k|\boldsymbol{x}_{lt}, \boldsymbol{y}_{lt}, \boldsymbol{\Theta})$ の計算において，収束条件に近い場合，次の近似が成り立つことを仮定する．

$$P(k|\boldsymbol{x}_{lt}, \boldsymbol{y}_{lt}, \boldsymbol{\Theta}^{(n)}) \approx P(k|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}^{(n)}) \tag{4.12}$$

この仮定により，E-step での計算にクリーン音声特徴量 $\boldsymbol{x}$ と $P(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Theta}, k)$ のモデルパラメータ $\boldsymbol{A}, \boldsymbol{\Sigma} \subset \boldsymbol{\Theta}$ が影響しなくなる．したがって，E-step の更新に関与するパラメータ $\boldsymbol{\mu}, \boldsymbol{\sigma}, \pi \subset \boldsymbol{\Theta}$ に関してのみ M-step で更新を行い収束した後に，収束後の $P(k|\boldsymbol{x}_{lt}, \boldsymbol{y}_{lt}, \boldsymbol{\Theta})$ を用いて $\boldsymbol{A}, \boldsymbol{\Sigma} \subset \boldsymbol{\Theta}$ を計算しても良いこととなる．つまり，あらかじめ観測音声特徴量 $\boldsymbol{y}$ の分布の GMM を学習しておき，それをパラメータ $\boldsymbol{\mu}, \boldsymbol{\sigma}, \pi \subset \boldsymbol{\Theta}$ の収束値として用いることと同値である．その場合，線形変換パラメータ $\boldsymbol{A}$ は式 (4.13) のように解析的に得ることができる．

$$\hat{\boldsymbol{A}}_k = \boldsymbol{X}\boldsymbol{R}_k\boldsymbol{Y}'^{\top}(\boldsymbol{Y}'\boldsymbol{R}_k\boldsymbol{Y}'^{\top})^{-1} \tag{4.13}$$

ここで，$\boldsymbol{X}$ と $\boldsymbol{Y}'$ はそれぞれパラレル学習データのクリーン音声特徴量と劣化音声特徴量を並べたものであり，$\boldsymbol{R}_k$ は対角成分に $[P(k|\boldsymbol{y}_1), P(k|\boldsymbol{y}_2), ..., P(k|\boldsymbol{y}_L)]$ を持つ対角行列となる．

評価時では，式 (4.3) を計算する．ここで，$\boldsymbol{x}$ は観測することができないため，直接に計算することは不可能である．そこで，評価時も EM-algorithm によって求めることを考える．$\boldsymbol{x}$ をパラメータの一つであると考えると，Q 関数は次式のようになる．

$$Q(\boldsymbol{x}_{lt}|\boldsymbol{x}_{lt}^{(n)}) = \mathrm{E}_{k|\boldsymbol{x}_{lt}^{(n)}, \boldsymbol{y}_{lt}, \boldsymbol{\Theta}}[\log P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}, k) P(\boldsymbol{y}_{lt}|\boldsymbol{\Theta}, k) P(k|\boldsymbol{\Theta})] \tag{4.14}$$

したがって，E-step と M-step を学習時と同様に表すと
E-step:

$$P(k|\boldsymbol{x}_{lt}^{(n)}, \boldsymbol{y}_{lt}, \boldsymbol{\Theta}) = \frac{P(\boldsymbol{x}_{lt}^{(n)}|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}, k) P(\boldsymbol{y}_{lt}|\boldsymbol{\Theta}, k) P(k|\boldsymbol{\Theta})}{\sum_{k'} P(\boldsymbol{x}_{lt}^{(n)}|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}, k') P(\boldsymbol{y}_{lt}|\boldsymbol{\Theta}, k') P(k'|\boldsymbol{\Theta})} \tag{4.15}$$

M-step:

$$\begin{aligned}\boldsymbol{x}_{lt}^{(n+1)} &= \underset{\boldsymbol{x}_{lt}}{\operatorname{argmax}}\, Q(\boldsymbol{x}_{lt}|\boldsymbol{x}_{lt}^{(n)}) \\ &= \underset{\boldsymbol{x}_{lt}}{\operatorname{argmax}} \sum_k P(k|\boldsymbol{x}_{lt}^{(n)}, \boldsymbol{y}_{lt}, \boldsymbol{\Theta}) \log P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}, k) P(\boldsymbol{y}_{lt}|\boldsymbol{\Theta}, k) P(k|\boldsymbol{\Theta})\end{aligned} \tag{4.16}$$

ここで同様に

$$P(k|\boldsymbol{x}_{lt}^{(n)}, \boldsymbol{y}_{lt}, \boldsymbol{\Theta}) \approx P(k|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}) \tag{4.17}$$

の近似が成り立つことを仮定すると，E-step の更新に $\boldsymbol{x}$ 自体が関与しなくなる．そのため，収束後の推定されるクリーン音声特徴量 $\hat{\boldsymbol{x}}_{lt}$ は

$$\hat{\boldsymbol{x}}_{lt} = \underset{\boldsymbol{x}}{\operatorname{argmax}} \sum_k P(k|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}) \log P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}, k) P(\boldsymbol{y}_{lt}|\boldsymbol{\Theta}, k) P(k|\boldsymbol{\Theta}) \tag{4.18}$$

$$= \underset{\boldsymbol{x}}{\operatorname{argmax}} \sum_k P(k|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}) \log P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}, k) \tag{4.19}$$

$$= \sum_k P(k|\boldsymbol{y}_{lt}) \boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y}_{lt} \end{bmatrix} \tag{4.20}$$

$$P(k|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}) = \frac{P(\boldsymbol{y}_{lt}|\boldsymbol{\Theta}, k) P(k|\boldsymbol{\Theta})}{\sum_{k'} P(\boldsymbol{y}_{lt}|\boldsymbol{\Theta}, k') P(k'|\boldsymbol{\Theta})} \tag{4.21}$$

として求めることができる．以上が SPLICE の定式化である．

さて，SPLICE において重要であるのは式 (4.2) である．つまり，局所的な入力特徴量空間で見た場合，入力である観測音声特徴量 $\boldsymbol{y}$ と出力であるクリーン音声特徴量 $\boldsymbol{x}$ の関係が線形変換の形で表すことが可能であるという点である．言い換えれば，局所的にでも線形変換で表すことが可能でさえあれば，音声特徴量だけに留まらず区分的線形変換は様々な利用が可能である．

4.2.3　実験

i)　実験条件

AURORA-2[31] を用いて音声認識実験により SPLICE の有効性を示す．AURORA-2 のタスクは英語連続数字発声であり，背景雑音が重畳されている．表 (4.1) に AURORA-2 のデータセットを示す．音響モデルは word 型 HMM を用い，クリーン音声のみで HMM を学習した．特徴量は MFCC とその対数パワー，そしてその 1 次微分と 2 次微分の計 39 次元を用い，全ての場合において CMN(Cepstral Mean Normalization) をかけている．SPLICE の GMM の混合数は 1024 を採用しており，分散は対角共分散行列を仮定して学習している．特徴量強調を行わない場合と SPLICE を行った場合での認識率を比較した．

表 4.1: AURORA-2 データセット

	training data	test set A	test set B
data	male 55, female 55	male 52, female 52	male 52, female 52
noise	subway, babble, car noise, exhibition hall	subway, babble, car noise, exhibition hall	restaurant, street, airport, train station
SNR	5dB, 10dB, 15dB, 20dB, clean	-5dB, 0dB, 5dB, 10dB, 15dB, 20dB, clean	-5dB, 0dB, 5dB, 10dB, 15dB, 20dB, clean

表 4.2: AURORA-2 認識結果 % (clean condition)

	no enhancement		SPLICE	
	test set A	test set B	test set A	test set B
CLEAN	99.37	99.37	99.40	99.40
SNR20	97.25	97.94	98.86	98.88
SNR 15	92.03	94.16	97.82	98.28
SNR 10	71.82	79.47	94.50	95.14
SNR 5	35.62	45.77	81.92	84.32
SNR 0	18.79	22.73	51.13	53.91
SNR -5	10.78	12.37	19.83	22.00

ii)　結果

特徴量強調を行わなかった場合と，SPLICE により特徴量強調を行った場合を比較した結果を表 (4.2) に示す．

iii)　考察

特に高雑音環境下で SPICE を行った場合非常に良い結果が出ている．これは，SPLICE により劣化音声特徴量からクリーン音声特徴量が推定できていることが示されている．ただし，SNR-5 の場合は学習データセットとのミスマッチが生じるため，認識率はそれほど上昇しない．次節では SPLICE の応用として雑音特徴量を用いた劣化音声特徴量変換について紹介する．

4.3　雑音特徴量を用いた劣化音声特徴量変換

区分的線形変換は先に述べた通り，局所的に線形変換で近似することが可能である場合，ドメインに関わらず応用することが可能である．この性質を利用した先行研究として観測特徴量と雑音の推定値の連結ベクトルからクリーン音声特徴量を推定する手法がある [32]．

雑音特徴量を $\boldsymbol{n}$ とすると，推定された雑音特徴量 $\hat{\boldsymbol{n}}$ を用いて

$$\hat{\boldsymbol{x}}_{lt} = \sum_k P(k|\boldsymbol{L} \begin{bmatrix} \boldsymbol{y}_{lt} \\ \hat{\boldsymbol{n}}_{lt} \end{bmatrix}) \boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y}_{lt} \\ \hat{\boldsymbol{n}}_{lt} \end{bmatrix} \tag{4.22}$$

として観測音声特徴量と雑音の連結ベクトルとクリーン音声特徴量の関係を線形変換で表現可能であるとしてモデル化を行っている．ここで，$\boldsymbol{L}$ は LDA による次元圧縮行列であり，学習にはあらかじめクリーン音声特徴量 $\boldsymbol{x}$ で学習した GMM を用いて計算される $P(m|\boldsymbol{x})$ をラベルとして用いる．このように入力と出力の次元数や意味が異なっていても適用可能である点は非常に特徴的である．

4.4　まとめ

本章では，区分的線形変換を用いた特徴量強調手法である SPLICE の説明を行い，その特性と拡張性について議論を行い，提案手法について定式化を行った．本提案手法の新規性は，区分的線形変換の枠組みにより特徴量強調に画像特徴量を利用することで，認識に直接画像情報を利用するだけでなく，様々な応用が可能になる点である．

次章より，提案手法における近似式 (5.11) の妥当性と，既存手法との比較により提案手法の有効性を実験的に示す．

第5章

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

5.1　はじめに

前章では区分的線形変換を用いた特徴量強調手法である SPLICE について取り上げ，その定式化と理論的意味について紹介した．加えて，その特性を説明し，拡張としていくつかの手法を取り上げた．SPLICE の最大の特徴は，VTS-based の特徴量強調のようにクリーン音声特徴量と観測音声特徴量と雑音特徴量の関係を明示的に持つわけではなく，線形変換で近似を行うことである．そのため，例えば画像情報のようにクリーン音声特徴量や観測音声特徴量との関係が不明瞭である特徴を自然な流れで導入することが可能となる．

本提案手法では，画像情報を区分的線形変換の枠組みにて特徴量強調に導入し，得られたクリーン音声特徴量の推定値を用いて音声認識を行うことで，雑音環境下での認識率の向上を計る．本章では，まず，提案手法の定式化と意味について説明をし，高速化のための近似についても議論を行う．

5.2　特徴量強調の枠組みによる画像情報と音声情報の統合

クリーンな環境において，画像情報のみによる音声認識は，人間は同じ口の形でも異なる音素を発声することができることからも分かる通り，音声情報のみによる音声認識に大きく劣る．そこで，本来，マルチモーダル音声認識において重要であるのは，音響雑音が小さな環境では音声情報を重視し，音響雑音が大きな環境では画像情報を重視することであると考えられる．マルチモーダル音声認識は，特徴量ベースで統合する初期統合と尤度ベースで統合する結果統合に大きく分けることができる．しかし，初期統合と結果統合は統合法が違えど，最終的な統合の際のそれぞれのストリームに対する重みをどのように決定するのかが問題であった．そこで，これを解決するために，区分的線形変換の枠組みを用いた初期統合法を導入する．局所的な領域で考えた際，クリーン音声特徴量を画像特徴量と音声特徴量の線形変換で近似することができると仮定をおくことで推定し，これを利用して音声認識を行う．これは，画像特徴量を補助として利用し，クリーン音声を推定する特徴量強調であると同時に，観測音声特徴量と画像特徴量から音声特徴量への次元圧縮を行う初期統合型のマルチモーダル音声認識の一種とも考えることができる．

図 (5.1) に提案手法の概形を示す．まず，それぞれのストリームで特徴量抽出を行う．その後，それらを結合したベクトルからクリーン音声特徴量の推定を行う．最終的に，得られたクリーン音声特徴量の推定値を用いて認識を行う．

5.3　区分的線形変換を用いたマルチモーダル特徴量強調

画像情報が特に高雑音環境下での音声認識の助けになることは自明と言えるが，どの段階で音声と画像の情報を統合するかがマルチモーダル音声認識の大きな課題の 1 つであった．本提案手法は区分的線形変換の，入力と出力間の制約の少なさという性質を利用して特徴量強調に口唇動画像から得られる画像特徴量を利用する．本提案手法の大きな利点として，クリーン音声特徴量を推定するため，単に音声認識に利用するだけではなく，音声

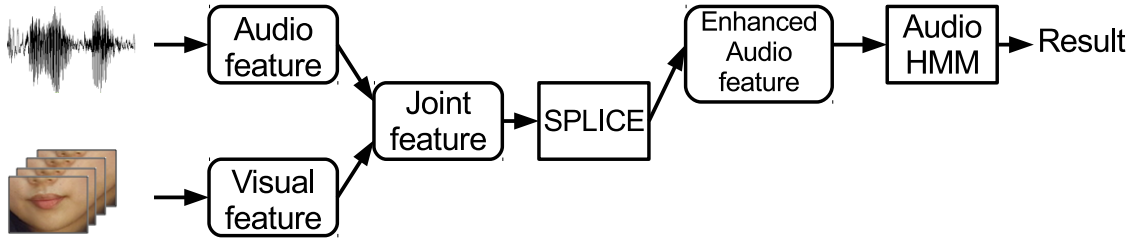

図 5.1: 提案手法のフロー図

再合成など様々な応用が可能であることが考えられる．まず提案手法の定式化を行い，その後近似についての検討を行う．

5.3.1　定式化

観測画像特徴量を $\boldsymbol{i}_{lt}$ とすると

$$\begin{aligned}\hat{\boldsymbol{x}}_{lt} &= \underset{\boldsymbol{x}_{lt}}{\operatorname{argmax}} \log \sum_k P(\boldsymbol{x}_{lt}, k|\boldsymbol{y}_{lt}, \boldsymbol{i}_{lt}, \boldsymbol{\Theta}) \\ &= \underset{\boldsymbol{x}_{lt}}{\operatorname{argmax}} \log \sum_k P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt}, \boldsymbol{i}_{lt}, \boldsymbol{\Theta}, k) P(\boldsymbol{y}_{lt}, \boldsymbol{i}_{lt}|\boldsymbol{\Theta}, k) P(k|\boldsymbol{\Theta}) \end{aligned} \tag{5.1}$$

を最大化することを考える．$P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt}, \boldsymbol{i}_{lt}, \boldsymbol{\Theta}, k)$ を

$$P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt}, \boldsymbol{i}_{lt}, \boldsymbol{\Theta}, k) = \mathcal{N}(\boldsymbol{x}_{lt}; \boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y}_{lt} \\ \boldsymbol{i}_{lt} \end{bmatrix}, \boldsymbol{\Sigma}_k) \tag{5.2}$$

としてクリーン音声特徴量の分布を観測音声特徴量と観測画像特徴量の連結ベクトルからの線形変換を平均に持つガウス分布であると仮定することにより，

$$\hat{\boldsymbol{x}}_{lt} = \underset{\boldsymbol{x}_{lt}}{\operatorname{argmax}} \sum_k P(k|\boldsymbol{x}_{lt}, \boldsymbol{y}_{lt}, \boldsymbol{i}, \boldsymbol{\Theta}) \log P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt}, \boldsymbol{i}_{lt}, \boldsymbol{\Theta}, k) \tag{5.3}$$

$$= \sum_k P(k|\boldsymbol{x}_{lt}, \boldsymbol{y}_{lt}, \boldsymbol{i}_{lt}) \boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y}_{lt} \\ \boldsymbol{i}_{lt} \end{bmatrix} \tag{5.4}$$

となる．これは同様に学習時と認識時共に EM-algorithm を用いて計算することが可能である．学習時における Q 関数は

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)}) = \mathrm{E}_{k|\boldsymbol{x},\boldsymbol{y},\boldsymbol{i},\boldsymbol{\Theta}^{(n)}}[\log P(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{i}, \boldsymbol{\Theta}, k) P(\boldsymbol{y}, \boldsymbol{i}|\boldsymbol{\Theta}, k) P(k|\boldsymbol{\Theta})] \tag{5.5}$$

となる．したがって E-step と M-step の計算は

E-step:

$$P(k|\boldsymbol{x}_{lt},\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta}^{(n)}) = \frac{P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta}^{(n)},k)P(\boldsymbol{y}_{lt},\boldsymbol{i}_{lt}|\boldsymbol{\Theta}^{(n)},k)P(k|\boldsymbol{\Theta}^{(n)})}{\sum_{k'}P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta}^{(n)},k')P(\boldsymbol{y}_{lt},\boldsymbol{i}_{lt}|\boldsymbol{\Theta}^{(n)},k')P(k'|\boldsymbol{\Theta}^{(n)})} \tag{5.6}$$

M-step:

$$\begin{aligned}\boldsymbol{\Theta}^{(n+1)} &= \operatorname*{argmax}_{\boldsymbol{\Theta}}\sum_l\sum_t Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)}) \\ &= \operatorname*{argmax}_{\boldsymbol{\Theta}}\sum_l\sum_t\sum_k P(k|\boldsymbol{x}_{lt},\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta}^{(n)})\log P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta},k)P(\boldsymbol{y}_{lt},\boldsymbol{i}_{lt}|\boldsymbol{\Theta},k)P(k|\boldsymbol{\Theta})\end{aligned} \tag{5.7}$$

となり，認識時の Q 関数は

$$Q(\boldsymbol{x}_{lt}|\boldsymbol{x}_{lt}^{(n)}) = \mathrm{E}_{k|\boldsymbol{x}_{lt}^{(n)},\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta}}[\log P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta},k)P(\boldsymbol{y}_{lt},\boldsymbol{i}_{lt}|\boldsymbol{\Theta},k)P(k|\boldsymbol{\Theta})] \tag{5.8}$$

したがって，E-step と M-step を学習時と同様に表すと
E-step:

$$P(k|\boldsymbol{x}_{lt}^{(n)},\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta}) = \frac{P(\boldsymbol{x}_{lt}^{(n)}|\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta},k)P(\boldsymbol{y}_{lt},\boldsymbol{i}_{lt}|\boldsymbol{\Theta},k)P(k|\boldsymbol{\Theta})}{\sum_{k'}P(\boldsymbol{x}_{lt}^{(n)}|\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta},k')P(\boldsymbol{y}_{lt},\boldsymbol{i}_{lt}|\boldsymbol{\Theta},k')P(k'|\boldsymbol{\Theta})} \tag{5.9}$$

M-step:

$$\begin{aligned}\boldsymbol{x}_{lt}^{(n+1)} &= \operatorname*{argmax}_{\boldsymbol{x}_{lt}} Q(\boldsymbol{x}_{lt}|\boldsymbol{x}_{lt}^{(n)}) \\ &= \operatorname*{argmax}_{\boldsymbol{x}_{lt}}\sum_k P(k|\boldsymbol{x}_{lt}^{(n)},\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta})\log P(\boldsymbol{x}_{lt}|\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta},k)P(\boldsymbol{y}_{lt},\boldsymbol{i}_{lt}|\boldsymbol{\Theta},k)P(k|\boldsymbol{\Theta})\end{aligned} \tag{5.10}$$

となる．

5.3.2　近似の導入

さて，ここで計算量の削減の点から SPLICE と同様に近似を導入する．SPLICE の場合と異なり，$P(k|\boldsymbol{x}_{lt},\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta})$ の近似の入れ方に大きく分けて 2 通りの手法が考えられる．1 つめは，SPLICE と同様に観測音声特徴量のみに依存することを仮定する．

$$P(k|\boldsymbol{x}_{lt},\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta}) \approx P(k|\boldsymbol{y}_{lt},\boldsymbol{\Theta}) \tag{5.11}$$

これは，SPLICE の結果からも分かる通り，雑音環境下でも妥当な近似であると考えられるが，異なる音素を発声していた場合でも雑音の影響で同じクラスタに分類されることは十分考えられる．２つめは，観測音声特徴量と観測画像特徴量の両方に依存する仮定をする．

$$P(k|\boldsymbol{x}_{lt},\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta}) \approx P(k|\boldsymbol{y}_{lt},\boldsymbol{i}_{lt},\boldsymbol{\Theta}) \tag{5.12}$$

これは，特徴量空間における局所領域のクラスタリングが画像特徴量の影響を受けることとなる．そのため，上記の雑音の影響で同じクラスタに分類される場合を回避できる可能性が生じる．しかし，この仮定は，口の形が同じであっても発声されるのは異なる音素であることは比較的頻発すると考えられるため，画像特徴量の強過ぎる影響で，適切なクラスタリングがなされない可能性も考えられる．そのため，本提案手法では式 (5.11) の近似を採用する．

この近似の導入により，モデルパラメータの学習時の Q 関数はこれを特徴量分布のパラメータに着目すると

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)}) = \mathrm{E}_{k|\boldsymbol{y},\boldsymbol{\Theta}^{(n)}}[\log P(\boldsymbol{y}|\boldsymbol{\Theta},k)P(k|\boldsymbol{\Theta})] \tag{5.13}$$

となる．これは，GMM 学習の際の式と同一であり，これによりあらかじめ GMM 学習を観測特徴量 $\boldsymbol{y}$ について行うことで特徴量分布のパラメータを得ることができる．その後，得られた特徴量分布のモデルパラメータを利用して $P(k|\boldsymbol{y}_{lt})$ を求める．

$$\begin{aligned} P(\boldsymbol{y}_{lt}|k) &= \mathcal{N}(\boldsymbol{y}_{lt};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \\ P(\boldsymbol{y}_{lt}) &= \sum_k \pi_k N(\boldsymbol{y}_{lt};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \\ P(k|\boldsymbol{y}_{lt}) &= \frac{\pi_k \mathcal{N}(\boldsymbol{y}_{lt};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{y}_{lt};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)} \end{aligned} \tag{5.14}$$

線形変換行列 $\boldsymbol{A}_k$ の学習は得られた $P(k|\boldsymbol{y}_{lt})$ を用いて

$$\{\hat{\boldsymbol{A}}_k\} = \underset{\{\boldsymbol{A}_k\}}{\operatorname{argmin}} \sum_l \sum_t \sum_k P(k|\boldsymbol{y}_{lt}) \left\| \boldsymbol{x}_{lt} - \boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y}_{lt} \\ \boldsymbol{i}_{lt} \end{bmatrix} \right\|^2 \tag{5.15}$$

として求めることができる．これは SPLICE と同様に解析解を得ることが可能であり，

$$\boldsymbol{A_k} = \boldsymbol{X}\boldsymbol{R}_k\bar{\boldsymbol{M}}^\top(\bar{\boldsymbol{Y}}\boldsymbol{R}_k\bar{\boldsymbol{M}}^\top)^{-1} \tag{5.16}$$

となる．ただし，$\boldsymbol{X}$ と $\bar{\boldsymbol{M}}$ はそれぞれクリーン音声と観測音声の特徴量 $[1, \boldsymbol{y}_{lt}^\top, \boldsymbol{i}_{lt}^\top]^\top$ を並べた物であり，$\boldsymbol{R}_k$ は対角成分に $[P(k|\boldsymbol{y}_1), P(k|\boldsymbol{y}_2), \ldots, P(k|\boldsymbol{y}_{LT})]$ を持つ対角行列を並べたものとなる．最終的なクリーン音声特徴量への変換は

$$\hat{\boldsymbol{x}}_{lt} = \sum_k P(k|\boldsymbol{y}_{lt})\boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y}_{lt} \\ \boldsymbol{i}_{lt} \end{bmatrix} \tag{5.17}$$

として推定することが可能となる．

5.4　まとめ

本章では，提案手法である区分的線形変換を用いたマルチモーダル音声認識について述べた．提案手法は，従来のマルチモーダル音声認識における初期統合法を拡張し，区分的

線形変換の枠組みによる特徴量強調に画像情報を利用することで，音声と画像のストリームを適切に統合することが可能となる．また，計算量の問題と特徴量空間のクラスタリング精度の観点から近似を導入した．次章では，近似の妥当性と本提案手法の有効性を音声認識実験により示す．

第6章

実験

6.1　はじめに

前章で述べた提案手法の有効性を実験により示す．コーパスとして CENSREC-1-AV を用いた認識実験を行った．これは，日本語連続数字読み上げタスクにおいて画像と音声の対応が取れたデータが収録されている．本章では，まず，5.3.2 節で述べた，提案手法における空間分割の近似の妥当性を実験により示す．その後，既存手法との比較を行い，本提案手法の有効性を示す．

6.2　近似の妥当性

本節では，5.3.2 節で述べた，提案手法における空間分割の近似の妥当性を実験により示す．本提案手法では，観測特徴量が得られた際の，各コンポーネントに対する重みを計算する際，

$$P(k|\boldsymbol{x}_{lt}, \boldsymbol{y}_{lt}, \boldsymbol{i}_{lt}, \boldsymbol{\Theta}) \approx P(k|\boldsymbol{y}_{lt}, \boldsymbol{\Theta}) \tag{6.1}$$

の近似を導入しており，これを用いた際の特徴量強調の式は，

$$\hat{\boldsymbol{x}}_{lt} = \sum_k P(k|\boldsymbol{y}_{lt}) \boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y}_{lt} \\ \boldsymbol{i}_{lt} \end{bmatrix} \tag{6.2}$$

となる．これは，コンポーネント重みが観測音声特徴量のみに依存し，画像情報に依存しないという仮定を行っている．実験により，画像情報に依存させてコンポーネント重みを決定する場合との比較を行う．画像情報に依存させてコンポーネント重みを決定する場合，近似と特徴量変換の式は

$$P(k|\boldsymbol{x}_{lt}, \boldsymbol{y}_{lt}, \boldsymbol{i}_{lt}, \boldsymbol{\Theta}) \approx P(k|\boldsymbol{y}_{lt}, \boldsymbol{i}_{lt}, \boldsymbol{\Theta}) \tag{6.3}$$

$$\hat{\boldsymbol{x}}_{lt} = \sum_k P(k| \begin{bmatrix} \boldsymbol{y}_{lt} \\ \boldsymbol{i}_{lt} \end{bmatrix}) \boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y}_{lt} \\ \boldsymbol{i}_{lt} \end{bmatrix} \tag{6.4}$$

となる．しかし，本提案手法の枠組みでは，計算量の削減のため GMM の共分散行列を対角行列として学習している．そのため，マルチモーダルの前提として相関があると考えている $\boldsymbol{y}_{lt}$ と $\boldsymbol{i}_{lt}$ の関係を無視していることに相当してしまい，これが問題となる．そこで，この問題を軽減するため，GMM の学習を音声と画像の連結ベクトルのみを PCA により 39 次元に次元圧縮したベクトル $\boldsymbol{z}_{lt}$ を用いて行う．

$$\boldsymbol{z}_{lt} = \boldsymbol{D} \begin{bmatrix} \boldsymbol{y}_{lt} \\ \boldsymbol{i}_{lt} \end{bmatrix} \tag{6.5}$$

ここで，$\boldsymbol{D}$ は PCA によって学習された変換行列である．これを用いて，特徴量強調を

$$\hat{\boldsymbol{x}}_{lt} = \sum_k P(k|\boldsymbol{z}_{lt}) \boldsymbol{A}_k^{PCA} \begin{bmatrix} 1 \\ \boldsymbol{y}_{lt} \\ \boldsymbol{i}_{lt} \end{bmatrix} \tag{6.6}$$

として行う．計算は，$\boldsymbol{z}_{lt}$ の確率密度関数を GMM と仮定して $P(k|\boldsymbol{z}_{lt})$ を求める．

$$\begin{aligned} P(\boldsymbol{z}_{lt}|k) &= \mathcal{N}(\boldsymbol{z}_{lt}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ P(\boldsymbol{z}_{lt}) &= \sum_k \pi_k N(\boldsymbol{z}_{lt}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ P(k|\boldsymbol{z}_{lt}) &= \frac{\pi_k \mathcal{N}(\boldsymbol{z}_{lt}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{z}_{lt}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \end{aligned} \tag{6.7}$$

線形変換 $\boldsymbol{A}_k^{PCA}$ の学習は提案手法と同様に連結ベクトル $[\boldsymbol{y}_{lt}^t op, \boldsymbol{i}_{lt}^t op]^t op$ とクリーン音声 $\boldsymbol{x}_{lt}$ のパラレルデータを用いて行った．

$$\{\hat{\boldsymbol{A}}_k^{PCA}\} = \underset{\{\boldsymbol{A}_k^{PCA}\}}{\operatorname{argmin}} \sum_l \sum_t \sum_k P(k|\boldsymbol{z}_{lt}) \left\| \boldsymbol{x}_{lt} - \boldsymbol{A}_k^{PCA} \begin{bmatrix} 1 \\ \boldsymbol{y}_{lt} \\ \boldsymbol{i}_{lt} \end{bmatrix} \right\|^2 \tag{6.8}$$

これを解くことにより線形変換 $\boldsymbol{A}_k^{PCA}$ は

$$\boldsymbol{A}_{\boldsymbol{k}}^{PCA} = \boldsymbol{X} \boldsymbol{R}_k^{PCA} \bar{\boldsymbol{M}}^\top (\bar{\boldsymbol{M}} \boldsymbol{R}_k^{PCA} \bar{\boldsymbol{M}}^\top)^{-1} \tag{6.9}$$

となる．ここで，$\boldsymbol{X}$ はクリーン音声の特徴量を並べた物，$\bar{\boldsymbol{M}}$ は $[1, \boldsymbol{y}_{lt}^\top, \boldsymbol{i}_{lt}^\top]^\top$ を並べた物であり，$\boldsymbol{R}_k^{PCA}$ は対角成分に $[P(k|\boldsymbol{z}_1), P(k|\boldsymbol{z}_2), \ldots, P(k|\boldsymbol{z}_{LT})]$ を持つ対角行列となる．PCA により空間を張り直すことで，GMM の対角以外の成分による影響を小さくすることができる．

6.2.1　実験条件

実験条件について説明を行う．コーパスには CENSREC-1-AV[33] を用いて実験を行った．コーパス中のデータの概要を表 (6.1,6.2) に示す．1 発話はそれぞれ 1 〜 7 桁の数字読み上げから構成されており，音声データと画像データは同期が取れているものとして扱う．画像に関しては雑音は想定せず，今回はクリーン条件でのみ学習・認識を行った．ここで SNR とは，音声対雑音比であり，単位は [dB] である．

表 (6.3) に実験に用いた特徴量を示す．音声特徴量は MFCC の 0 次元目を含めた 13 次元とそれの 1 次微分，2 次微分を用いる．本実験では，音声と画像のストリームの統合法についての検討を重視したため，画像特徴量は Appearance ベースのものを用いた．図 (6.1) に画像特徴量の抽出法を示す．まず，Raster scan によりベクトル化した後に，主成分分析により次元圧縮して得られた各色１０次元の合計３０次元を用いている．その後，音声と画像のフレームレートが違うため，画像特徴量を線形補完して音声とフレーム単位で同期さ

表 6.1: 音声データ （近似の妥当性の検証）

Sampling rate	16 kHz
Quantization bits	16 bit/sample
Audio noise	driving noises 2 types（cityroad，expressway） 9 SNR levels (-20dB to +20dB)

表 6.2: 画像データ （近似の妥当性の検証）

frame rate	29.97 Hz
pixel	24 bit color
data size	81 pixel width × 55 pixel height
Visual noise	none (clean)

表 6.3: 音声と画像特徴量 （近似の妥当性の検証）

Audio	MFCC+Δ+Δ^2 (39 dimensions)
Visual	Raster scan + PCA (10×3(RGB) = 30 dimensions)

表 6.4: データセット（近似の妥当性の検証）

	HMM	SPLICE	Test
speaker	male 22 female 20	male 22 female 20	male 25 female 26
audio data	clean	clean cityroad expressway	clean cityroad expressway
visual data	clean color (RGB)		

せた．今回の実験では，CENSREC-1-AV に収録されているデータを表 (6.4) のように認識用 HMM 学習セット，SPLICE 学習用パラレルデータセット，テストセットに分けて用いた．また，前述にように SPLICE の学習にはノイジー音声とクリーン音声のパラレルデータが必要だったため，CENSREC-1-AV の学習データにノイズを重畳することによって擬似的なパラレルデータを作成した．GMM の混合数は，1024 で行い，分散は対角共分散行列を仮定して行っている．

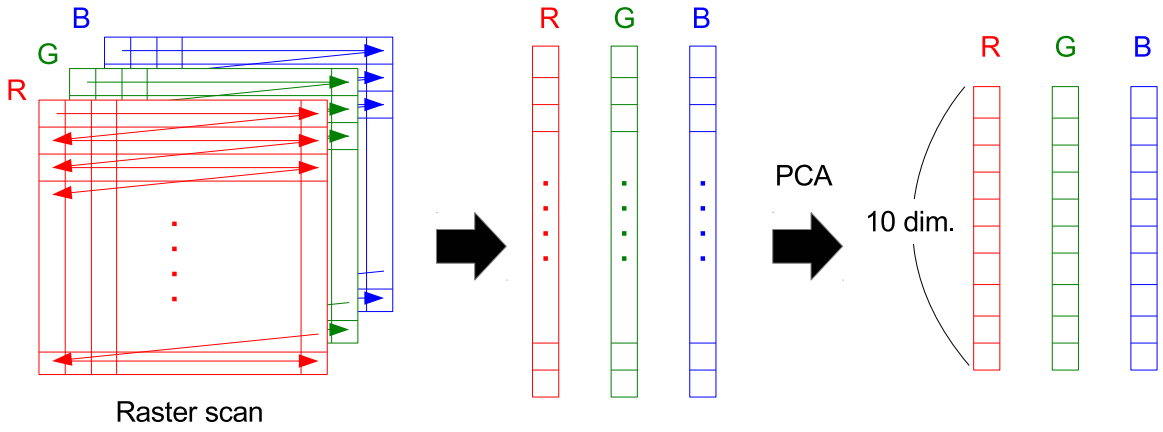

図 6.1: 画像特徴量

表 6.5: 近似の種類による認識率の比較 % (city-road noise)

SNR	Audio only	Audio & Visual	Audio & Visual with PCA
clean	**99.74**	99.72	99.71
20dB	**99.72**	99.71	**99.72**
10dB	**99.72**	**99.72**	99.71
0dB	**99.71**	99.60	99.63
-10dB	**98.87**	97.87	97.81
-20dB	**79.20**	71.01	75.41
average	**96.16**	94.61	95.11

表 6.6: 近似の種類による認識率の比較 % (city-road noise)

SNR	Audio only	Audio & Visual	Audio & Visual with PCA
clean	**99.74**	99.72	99.71
20dB	**99.71**	**99.71**	99.69
10dB	**99.71**	99.69	99.66
0dB	**99.64**	99.49	99.49
-10dB	**98.01**	96.59	96.94
-20dB	**71.70**	64.68	70.17
average	**94.75**	93.31	94.07

6.2.2　結果

提案手法である観測音声特徴量のみに依存させてコンポーネント重みを決定する式 (6.2) を画像情報にも依存させてコンポーネント重みを決定する式 (6.4) とさらに PCA により非対角成分の影響を低減した式 (6.6) を比較した．実験結果を表 (6.5,6.6) に示す．それぞれ，認識に用いる HMM をクリーン音声のみで学習したもの (clean conditon) と雑音が重畳された音声を含めたデータで学習したもの (multi condition) とで認識を行った結果である．

6.2.3　考察

ほぼ全ての場合において，提案手法である，コンポーネントの重み付けを観測音声特徴量のみに依存させる場合が最も良い結果が得られた．画像情報を利用した場合は，PCA の有無で大きな差が出ている．これは，対角共分散行列を仮定した学習を行っているため，PCA を行わない場合では音声と画像間の相関が無視されてしまっているためだと考えられる．また，提案手法である観測音声特徴量のみに依存してコンポーネント重みを決定する手法は，画像情報を用いた場合よりも良い結果が得られた．これは，PCA の変換に対する画像情報の寄与が大きかったために，最も GMM にて考慮されるべきノイズの大きさや種類が逆に考慮されなかったためだと考えられる．

次節では，本実験において最も結果の良かった観測音声特徴量のみに依存させてコンポーネントの重みを決定する式 (6.2) と従来手法との比較を行う．

6.3　従来手法との比較

本提案手法の有効性を従来手法との比較により示す．従来手法には，音声強調をかけないもの (no enhance)，初期統合 (feature fusion (PCA))，結果統合 (decision fusion) を用いた．また，画像を用いることによる効果を調べるために，音声特徴量のみを用いる SPLICE とも比較をおこなった．

6.3.1　実験条件

実験条件を示す．クリーン音声データは CENSREC-1-AV のものを用いるが，ノイジー音声データは雑音コーパスである noisex92[34] の雑音データを重畳することによって作成した．SNR は 20dB から 0dB までの 5 段階を採用した．初期統合は音声と画像の特徴量を連結した 69 次元のベクトルを PCA により 39 次元に次元圧縮を行っている．結果統合はマルチストリーム HMM を採用し，対数尤度を統合する際の重み λ を 0.0，0.2，0.4，0.6，0.8，1.0 の内から最も良い認識率が得られるものを各雑音，各 SNR ごとに選択している．また，画像を用いることによる効果を調べるために，音声特徴量のみを用いる SPLICE とも比較をおこなった．全ての場合において，CMN(Cepstral Mean Normalization) をかけている．また，全ての場合において，認識用の HMM は 39 次元の特徴量ベクトルで学習を行った．

6.3.2　結果

認識実験の結果を図 (6.2) に示す．これは，全ての雑音と SNR における単語誤り率（Word Error Rate ; WER）を平均したものである．clean 条件と multi 条件は，それぞれクリーン音声のみで学習した HMM と雑音も用いて学習した HMM による認識結果である．また，図 (6.3–6.7) にそれぞれの種類の雑音ごとに全ての SNR の結果を平均したものを示す．

6.3.3　考察

実験の結果，従来手法 (no enhancement, feature fusion, decision fusion) の中では最も結果統合が良い結果を得られた．提案手法は，結果統合よりもクリーン音声のみで学習した HMM で 25%，雑音も用いて学習した HMM で 24%のエラー削減率を得ることができた．また，SPLICE と比較した場合も提案手法が認識率が良い結果となった．これにより，区分的線形変換による特徴量強調である SPLICE に画像特徴量も用いることの効果を示すことができたと言える．

また，雑音の種類ごとの認識結果を見ても，ほぼ全ての条件で提案手法が最も良い結果を得ることができた．例外として，走行時雑音（car noise）における multi 条件では，結果統合が最も良い結果を示したが，これは，そもそも何も耐雑音処理をしない場合 (no enhancement) でも非常に単語誤り率が低いことが原因である．全ての雑音の SNR におい

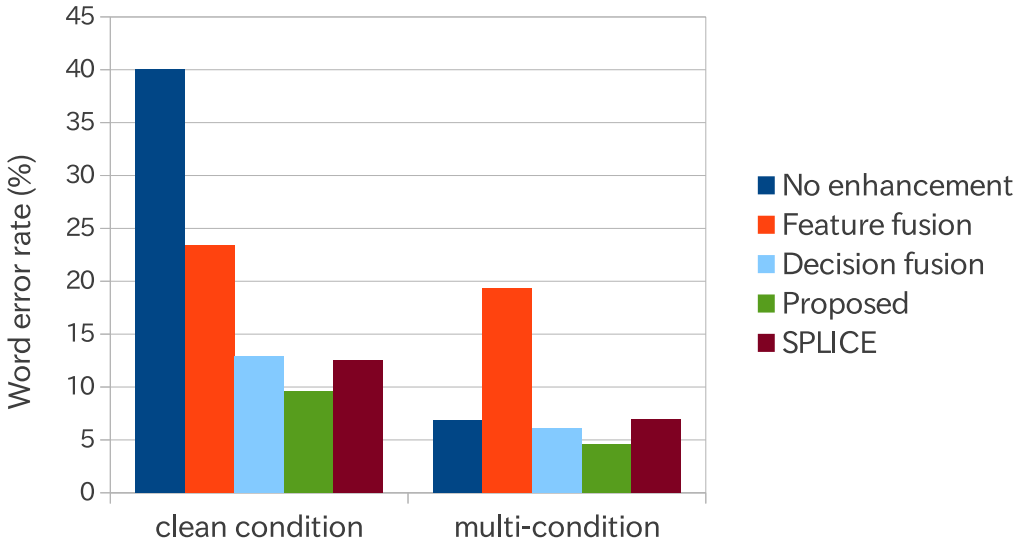

図 6.2: 単語認識誤り率（平均）

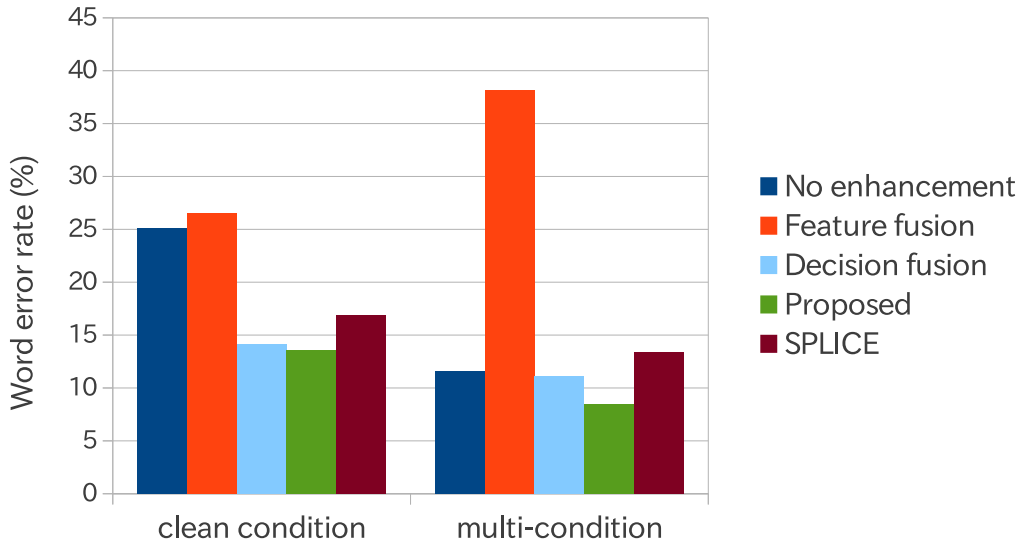

図 6.3: 単語認識誤り率（babble noise）

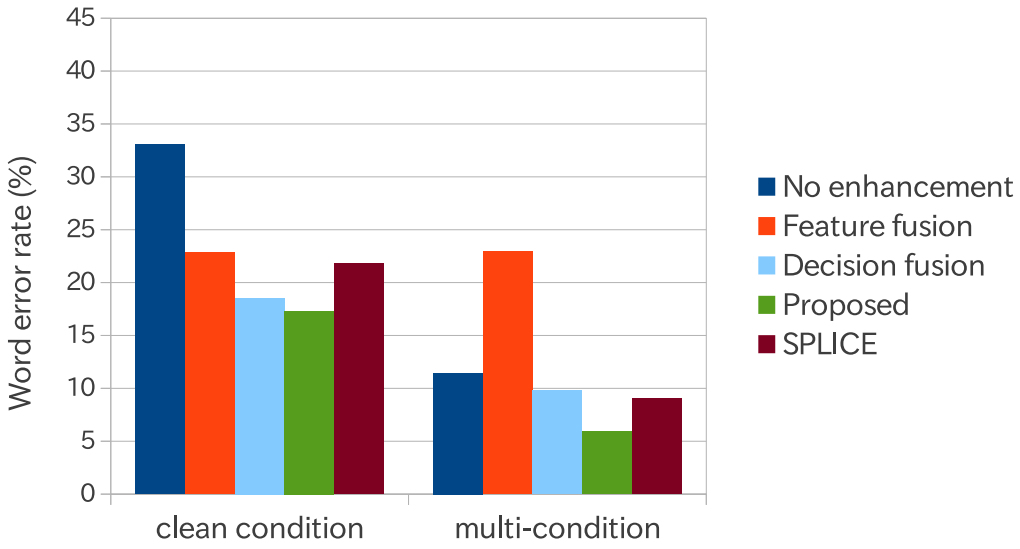

図 6.4: 単語認識誤り率（factory1 noise）

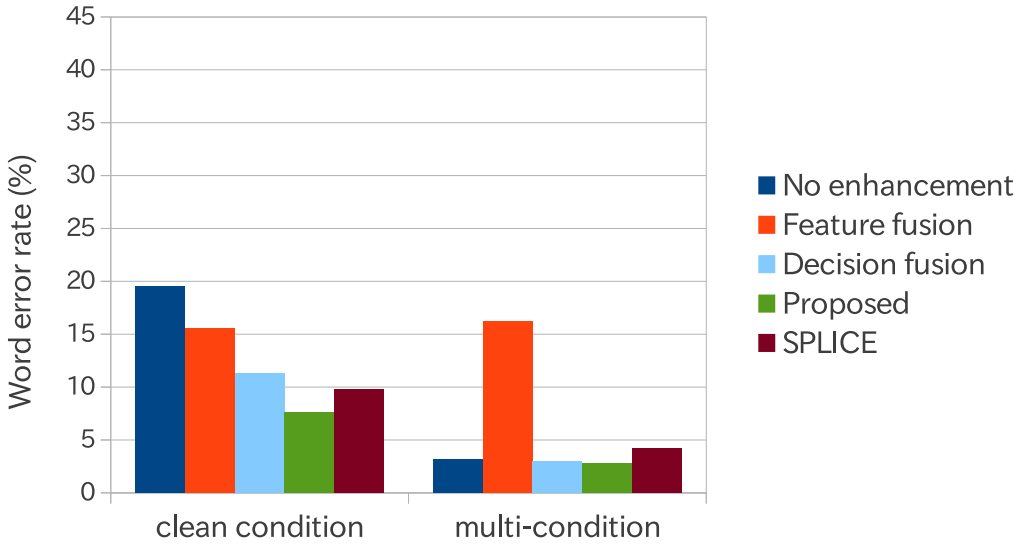

図 6.5: 単語認識誤り率（fatory2 noise）

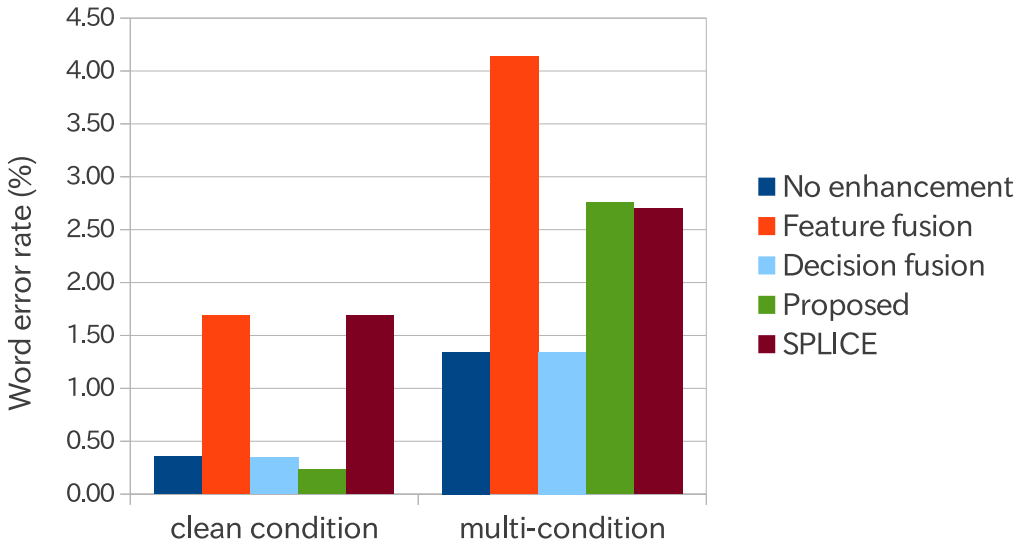

図 6.6: 単語認識誤り率（volvo (car) noise）

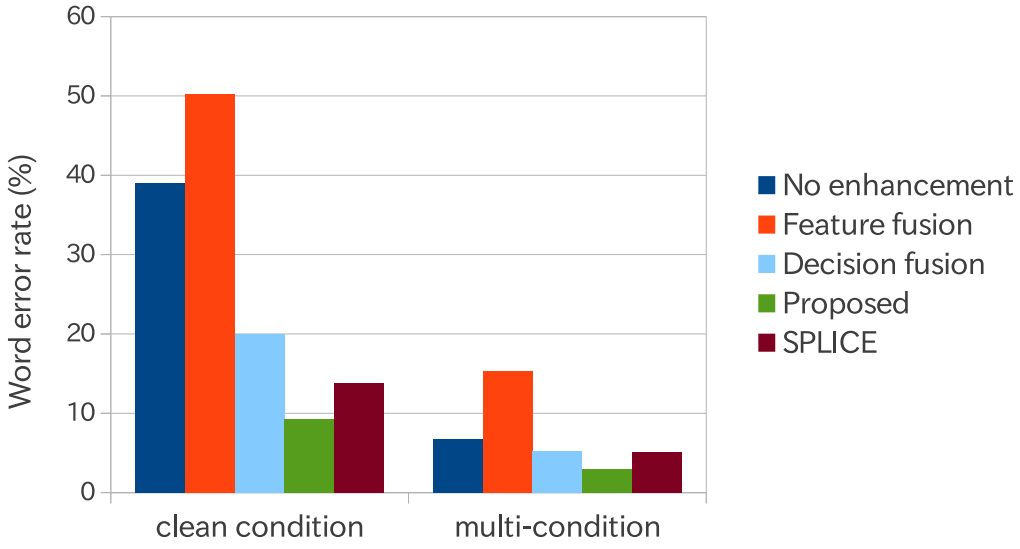

図 6.7: 単語認識誤り率（white noise）

表 6.7: 音声データ （従来手法との比較）

Samplling rate	16kHz
Auantization bits	16 bit/sample
Audio noise	noisyx-92 5 types (babble,factory1, factory2,car (Volvo),white) 5 SNR levels (20 dB to 0dB)

表 6.8: 音声データ （従来手法との比較）

Frame rate	29.97Hz
Pixsel	24 bit color
Data size	81 pixel width × 55 pixel height
Visual noise	none (clean)

表 6.9: 音声と画像特徴量（従来手法との比較）

Audio	MFCC+Δ+Δ^2
Visual	Raster scan + PCA (10 × 3 (RGB) = 30 dimensions)

表 6.10: データセット （従来手法との比較）

	HMM	Trans.	Test
Speaker	male 22 female 20		male 25 female 26
Audio data	clean noisyx-92 (babble, factory1, factory2, car (Volvo), white) 20dB 15dB 10dB 5dB 0dB		
visual data	clean color (RGB)		

ても，もともと単語誤り率が非常に低い箇所では同様の傾向が見られる．しかし，これは，相対的に見れば微小な劣化でありほとんど問題とならない．

6.4　まとめ

本章では，第 5 章で述べた提案手法の有効性を示すための実験とその結果について述べ，考察を行った．まず，提案手法における近似の妥当性を実験により検証した．CENSREC-1-AV を用いた認識実験によりコンポーネント重みを決定する際に観測音声特徴量にのみ依

存させた場合が最も良い結果が得られた．また，従来手法と比較を行い，従来手法で最も結果の良かったマルチストリーム HMM を用いたマルチモーダル手法に対して，CENSREC-1-AV を用いた認識実験により，提案手法は，従来手法で最も結果の良かったマルチストリーム HMM を用いた手法と比較し，クリーン音声のみで学習した HMM で 25%，雑音も用いて学習した HMM で 24%のエラー削減率を得ることができた．また，雑音音声からクリーン音声を区分的線形変換によって推定する SPLICE と比較した場合でも本提案手法は良い結果を得ることができた．これによって，実験により特徴量強調に画像情報を利用することの有効性が示された．

第7章

結論

7.1　本研究の成果

本研究は，口唇動画像を用いたマルチモーダル音声認識における新しいアプローチとして，特徴量強調の枠組みで音声情報と画像情報を統合する枠組みを提案した．雑音環境下における音声認識では，音声情報と画像情報を雑音や発話された音素などによって適切に組み合わせる必要があると考えられる．本研究の特色は，観測音声に依存させて音声と画像の統合法を適切に変更している点と，画像情報を利用してクリーン音声を推定しそれを用いて認識を行っている点である．クリーン音声への変換を観測音声に依存させているため，ヒトに近い認識プロセスが実現され，精度向上に有効である．また，画像情報を用いてクリーン音声を推定する枠組みであるため，音声認識応用の他にも，例えば音声再合成を行うことにより，高雑音環境下での通話品質の向上などが実現できる．

本手法の有効性を示すために，日本語数字読み上げに対する認識実験による性能評価を行った．性能評価には，日本語数字読み上げの口唇画像と音声の対応が取れているコーパスである CENSREC-1-AV を利用した．その結果，提案手法は，従来手法で最も結果の良かったマルチストリーム HMM を用いた手法と比較し，クリーン音声のみで学習した HMM で 25%，雑音も用いて学習した HMM で 24%のエラー削減率を得ることができた．以上より，本研究は新しいマルチモーダル音声認識の枠組みを提案することが出来たと言える．

7.2　今後の展望

今後の展望と課題についてマルチモーダル音声認識の各段階ごとに述べる．まず，特徴量に関してであるが，画像情報に雑音が乗った場合における検討を密にする必要がある．これは，単なるコントラストの変化等だけではなく，顔の正面画像が正確に得られない場合など様々なシチュエーションが考えられる．特に顔の向きが変化した場合は，音響的な特徴量も顔の向きに依存して変化すると考えられるため，検討が必要である．

また，モデル自体の改善としては，識別的な基準によりモデルパラメータを調整するなどが考えられる．今回は実験的にコンポーネントの重みを観測音声特徴量のみに依存して決定したが，基準の改善や，モデル混合数の増加などによってさらなる検討を行う必要が考えられる．

最後に，近年，多層パーセプトロンによる高精度な音響モデルが登場している．これは，特徴量強調にも応用することが可能であると考えられ，コンポーネント重みの決定に利用するなど応用先が多岐に渡ると予想される．そのため，今後は様々な特徴量やシチュエーションを網羅しつつ，モデルの改善を行うことが課題と考える．

謝辞

本研究ならびに本論文の執筆にあたり，多大なる御指導，御鞭撻を賜りました指導教員の広瀬啓吉教授ならびに峯松信明教授に深く感謝いたします．また，研究活動を様々な面で支えて下さった高橋登技官，秘書の池上恵さん，折茂結実子さんに深く感謝します．

また，本研究を進めるに当たって多大な助言をして頂いた博士課程の鈴木雅之氏には感謝の念が絶えません．鈴木雅之氏には研究の相談だけでなく，研究に対する姿勢等教わりました．

そして，音声コーパスを提供して頂きました各機関には大変お世話になりました．感謝致します．コーパスなくして本研究の結実はありませんでした．

また，インターンとして２ヶ月半程の間，NTT コミュニケーション科学研究所の皆様ならびに指導して頂いた久保陽太郎氏には大変お世話になりました．最前線で戦っている研究者と共に研究することができたインターンでの経験はかけがえのない物となったと思います．

日頃の研究室生活においても，研究に関して助言をして頂いた齋藤大輔助教，英論の言い回しの相談に乗って頂いたショートグレッグ氏，お互い博士進学を志し切磋琢磨し合った橋本浩弥氏，うどんに関する情報を提供して頂いた加藤集平氏，研究からサーバまで多くのことを相談させて頂いた甲斐常伸氏，良き遊び相手でもあった川口拓也氏には大変お世話になりました．その他大勢の研究室の皆様のお陰で，すばらしい修士課程を過ごせたと思います．ここに深く感謝の意を述べたいと思います．

最後に学生生活を支えて頂いた家族に感謝致します．有難うございました．

2013 年 2 月 6 日

柏木陽佑

参考文献

[1] Shen, Y.J. and Yang, S.P.: "A new Blind-Source-Separation method and its application to fault diagnosis of rolling bearing," International Journal of Nonlinear Sciences and Numerical Simulation, vol. 7, no. 3, pp. 245–250, 2011.

[2] Atal, BS.: "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," the Journal of the Acoustical Society of America, vol. 55, no. 6, pp. 1304–1312, 1974.

[3] Stouten, V.: "Robust automatic speech recognition in time-varying environments," KU Leuven, Diss, 2006.

[4] Droppo, J. and Deng, L. and Acero, A.: "Evaluation of the SPLICE on the Aurora2 and 3 Tasks," International Conference on Spoken Language Processing, pp. 29–32, 2002.

[5] Lee, C.H. and Lin, C.H. and Juang, B.H.: "A study on speaker adaptation of the parameters of continuous density hidden Markov models," Signal Processing, IEEE Transactions on, vol. 39, no. 4, pp. 806–814, 1991.

[6] Gauvain, J.L. and Lee, C.H.: "Maximum a posteriori estimation for multivariate gaus- sian mixture observations of Markov chains," Speech and Audio Processing, IEEE Transactions on, vol. 2, no. 2, pp. 291–298, 1994.

[7] Leggetter, CJ and Woodland, PC.: "Maximum likelihood speaker adaptation of continuous density hidden Markov models," Computer speech and language, vol. 9, no. 2, pp. 171, 1995.

[8] Ragni, A. and Gales, MJF.: "Structured discriminative models for noise robust continuous speech recognition," Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 4788–4791, 2011.

[9] Tariquzzaman, M. and Gyu, S.M. and Young, K.J. and You, N.S. and Rashid, MA.: "Performance Improvement of Audio-Visual Speech Recognition with Optimal Reliability Fusio," Internet Computing & Information Services (ICICIS), 2011 International Conference on, pp. 203–206, 2011.

[10] McGurk, H. and MacDonald, J.: “Hearing lips and seeing voices,” Nature, 1976.

[11] Nakamura, K. and Janke, M. and Wand, M. and Schultz, T.: “Estimation of fundamental frequency from surface electromyographic data: EMG-to-F¡ inf¿ 0¡/inf¿,” Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 573–576, 2011.

[12] Almajai, I. and Milner, B.: “Using audio-visual features for robust voice activity detection in clean and noisy speech,” Proc. EUSIPCO, vol. 7, pp. 19, 2008.

[13] Basu, S. and Neti, C. and Rajput, N. and Senior, A. and Subramaniam, L. and Verma, A.: “Audio-visual large vocabulary continuous speech recognition in the broadcast domain.,” Multimedia Signal Processing, 1999 IEEE 3rd Workshop on, pp. 475–481, 1999.

[14] Bregler, C. and Konig, Y.: “’Eigenlips’ for robust speech recognition,” Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on, vol. 2, pp. II–669, 1994.

[15] Dupont, S. and Luettin, J.: “Using the multi-stream approach for continuous audio-visual speech recognition,” 1997.

[16] Duchnowski, P. and Hunke, M. and Busching, D. and Meier, U. and Waibel, A.:“Toward movement invariant automatic lip-reading and speech recognition,” Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on, vol. 1, pp. 109–112, 1995.

[17] Potamianos, G. and Verma, A. and Neti, C. and Iyengar, G. and Basu, S.: “A cascade image transform for speaker independent automatic speechreading,” Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on, vol. 2, pp. 1097–1100, 2000.

[18] Chan, M.T. and Zhang, Y. and Huang, T.S.: “Real-time lip tracking and bimodal continuous speech recognition,” Multimedia Signal Processing, 1998 IEEE Second Workshop on, pp. 65–70, 1998.

[19] Potamianos, G. and Graf, H.P. and Cosatto, E.: “An image transform approach for HMM based automatic lipreading,” Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, pp. 173–177, 1998.

[20] Rogozan, A. and Deléglise, P. and Alissali, M.: “Adaptive determination of audio and visual weights for automatic speech recognition,” Audio-Visual Speech Processing: Computational & Cognitive Science Approaches, 1998.

[21] Basu, S. and Oliver, N. and Pentland, A.: “3D modeling and tracking of human lip motions,” Computer Vision, 1998. Sixth International Conference on, pp. 337–343, 1998.

[22] Chiou, G.I. and Hwang, J.N.: “Lipreading from color video,” Image Processing, IEEE Transactions on, vol.6, no. 8, pp. 1192–1195, 1997.

[23] Kass, M. and Witkin, A. and Terzopoulos, D.: “Snakes: Active contour models,” International journal of computer vision, vol.1, no. 4, pp. 321–331, 1988.

[24] Kaucic, R. and Dalton, B. and Blake, A.: “Real-time lip tracking for audio-visual speech recognition applications,” Computer Vision—ECCV’96, pp. 376–387, 1996.

[25] Cootes, T.F. and Edwards, G.J. and Taylor, C.J.: “Active appearance models,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 23, no. 6, pp. 681–685, 2001.

[26] Dupont, S. and Luettin, J.:“Audio-visual speech modeling for continuous speech recognition,” Multimedia, IEEE Transactions on, vol. 2, no. 3, pp.141-151, 2000.

[27] Potamianos, G. and Graf, H.P.: “Discriminative Training of HMM Stream Exponents for Audio- Visual Speech Recognition,” Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, vol. 6, pp. 3733–3736, 1998.

[28] Potamianos, G. and Neti, C. and Gravier, G. and Garg, A. and Senior, A.W.: “Recent advances in the automatic recognition of audiovisual speech,” Proceedings of the IEEE, vol. 91, no. 9, pp. 1306–1326, 2003.

[29] Potamianos, G. and Luettin, J. and Neti, C. “Hierarchical discriminant features for audio-visual LVCSR,” Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP’01). 2001 IEEE International Conference on, vol. 1, pp. 165–168, 2001.

[30] Hagen, A. and Morris, A.: “Recent advances in the multi-stream HMM/ANN hybrid approach to noise robust ASR,” Computer Speech & Language, vol. 19, no. 1, pp. 3–30, 2005.

[31] Hirsch, H.G. and Pearce, D.: “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW), 2000.

[32] Suzuki, M. and Yoshioka, T. and Watanabe, S. and Minematsu, N. and Hirose, K.: “MFCC enhancement using joint corrupted and noise feature space for highly non-stationary noise environments, ”Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference onl, pp. 4109–4112, 2012.

[33] Tamura, S. and Miyajima, C. and Kitaoka, N. and Yamada, T. and Tsuge, S. and Takiguchi, T. and Yamamoto, K. and Nishiura, T. and Nakayama, M. and Denda, Y. and Fujimoto, M. and Matsuda, S. and Ogawa, T. and Kuroiwa, S. and Takeda, K. and Nakamura, S.: "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," Proc. International Conference on Auditory-Visual Speech Processing, AVSP 2010, 2010.

[34] Varga, A. and Steeneken, H.J.M.: "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," vol. 12, no. 3, pp. 247–251, 1993.

発表文献

[1] 柏木陽佑, 鈴木雅之，峯松信明，広瀬啓吉: “SPLICE に基づく音声・口唇画像情報を用いた雑音環境下音声認識,” 日本音響学会春季研究発表会, 2012.

[2] 柏木陽佑, 鈴木雅之，峯松信明，広瀬啓吉: “区分的線形変換を用いた雑音環境下マルチモーダル音声認識,” 日本音響学会秋季講演論文集, pp.25–28, 2012.

[3] 柏木陽佑, 鈴木雅之，峯松信明，広瀬啓吉: “Audio-visual feature integration based on piecewise linear transformation for noise robust automatic speech recognition,” Proc. Spoken Language Technology (SLT), 2012.

[4] 柏木陽佑, 久保陽太郎, 峯松信明, 広瀬啓吉: “Deep Neural Network 混合モデルを用いた環境・話者適応の検討,” 日本音響学会春季研究発表会, 2013（発表予定）.