3 画像と音声情報を統合した発話認識

有木 康雄，駒井 祐人（神戸大学大学院システム情報学研究科）

唇・音声統合認識の現状

●統合認識が必要とされる背景

近年，音声認識はパソコンや携帯電話のソフトとして広まり，キーボードからの入力に替わるハンズフリーな技術として注目を集めている．しかし，現在の音声認識技術は，雑音環境下では認識性能が著しく低下してしまうという問題点があり，音声認識の環境適応に向けて大きな課題となっている．

一方，人間は発話内容を理解する際，種々の情報を総合的に利用している．たとえば，音声が聞き取りにくい場合，発話者の唇の動きに注目して発話内容を理解しようとし，逆に，唇の動きと音声が一致しない場合には，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．このように，人間による発話内容の理解には，唇の動き情報と音声情報の統合的利用がきわめて重要である．

唇の動きから発話内容を読み取る技術は読唇と呼ばれ，聴覚障害者のコミュニケーションの手段の1つとして用いられている．音声認識は雑音に弱く，雑音環境下では認識率が著しく低下するが，読唇は雑音に影響されることがない．そこで，雑音環境下で頑健に音声認識を行う手法として，唇の動き情報と音声情報を統合した唇・音声統合認識が注目され，近年研究が進められている．

●期待される応用例

音声認識の応用例として，音声によるカーナビゲーションや，音声対話受付案内システム，音声入力による文書作成や Web 検索，会議の議事録作成など，さまざまな研究が行われている．たとえば，音声入力による文書作成は，高齢者にとって複雑なキーボード操作を伴わないため，情報バリアフリーの点からも意義深い．唇・音声統合認識を用いれば，これらのアプリケーションの精度向上が期待できる．

さらに，監視カメラなど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も，発話内容の分析を行って，犯罪防止に役立てることなども期待できる．また，音声情報から顔の 3D アニメーションを生成する技術なども研究されている．相手の声に合わせて 3D アニメーションの顔が動くと，読唇による理解を可能とする☆1 ため，聴覚障害者にとってはコミュニケーションの手段として用いることができる．

実際に唇・音声統合認識を用いる際には，音声雑

☆1 読唇技術を身につけるには，相当の訓練が必要であり，聴覚障害者が皆この技術を身につけているわけではない．

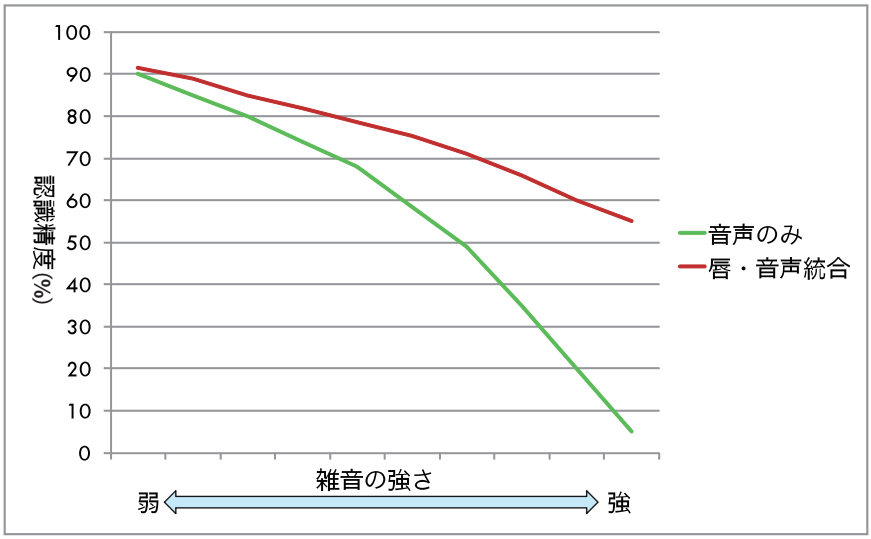

図-1　唇・音声統合認識の効果

音だけでなく，画像の雑音，すなわち，解像度の低下や，照明変動，顔方位のある発話などにも対応する必要がある．唇・音声統合認識をより実用化させるには，このような実環境で，認識精度を保証することが重要である．

●唇・音声統合認識の現状

現在の音声認識技術は，大規模なデータベースと統計理論に基づいており，大語彙連続音声認識や，対話音声認識に関する研究が行われている．認識精度も，丁寧に発話された音声であれば，90% 以上の精度が得られるようになっている．

一方，唇・音声統合認識では，単語や連続数字の発話認識に関する研究が多く，大語彙連続音声認識に関する研究は少ない．これは，通常の音声認識に比べ，唇の動き情報による認識精度が低いことから，小さなタスクで認識率を改善することが目的となっていること，また，視覚情報を含む大規模な日本語データベースが少ないことに起因している．

唇の動き情報のみで認識を行った場合，音声認識ほど高い精度を出すことができない．この理由は，音声では発声の基本単位である音素をある程度認識できるが，唇の動きからでは「か /ka/」と「が /ga/」といった子音の区別ができないからである．しかし，唇の動きと音声を統合することにより，雑音が重畳した場合に，認識率の低下を和らげるだけでなく，雑音が弱い状況下においても，認識精度を向上させる効果がある．図-1 は，英語の連続音声認識における研究成果を示したものである[1]．話者 239 人分の 17,000 発話を使って，後述する音素 HMM を学習させ，話者 26 人分の 1,000 発話を認識した場合の単語認識率を示している．単語の総数は，約 10,000 単語である．図を見ると，雑音が大きい状況下では，唇と音声を統合させることで，認識精度が大きく改善されていることが分かる．

唇・音声統合認識では近年，唇の動き情報と音声情報を統合した発話区間検出（VAD：Voice Activity Detection）が活発に研究されている．たとえば，ロボットに音声認識機能を搭載した場合，ロボットはいつ発話が行われるか分からないため，常に認識し続けることになり，雑音などに反応して誤認識が発生する．そこで，唇・音声統合により発話区間を検出し，その区間でのみ音声認識を行うようにすれば，誤認識を防ぐことができる．

唇と音声の認識技術

唇・音声統合認識を実現するには，まず，音声情報と唇の動き情報を抽出する必要がある．音声はマイクで入力できるが，唇は動画像のどの位置にあるのか分からないため，発話認識を行う際の前処理として，唇領域の抽出が必要である．その後，唇と音

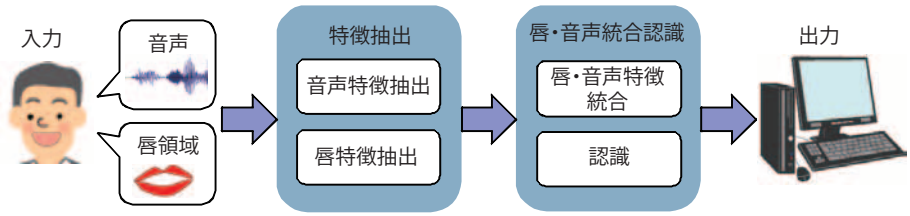

図-2　唇・音声統合認識の構成図

声の特徴量抽出を行い，唇・音声統合認識を行う．図-2に大まかな唇・音声統合認識の構成図を示す．

●唇領域抽出

唇領域の抽出法は，色情報を用いた抽出法や，あらかじめ用意しておいたテンプレートに近い部分を参照画像から探索するテンプレートマッチング，エネルギー関数が最小となるように制御点座標を更新するSnake，特徴点座標の系列に対してPCA（Principal Component Analysis：主成分分析）を行い，部分空間を構成して形状を表現するActive Shape Model（ASM），同様にアピアランスについても考慮するActive Appearance Model（AAM）などを用いたものがある．図-3にAAMによって抽出した唇領域の例を示す．AAMは初期探索点を探索領域から近い場所に与えれば，図-3のように特徴点座標を正確に抽出できる．

唇領域を抽出する際，図-3のように，唇領域だけでなく，顔の下側を大まかに抽出することが多い．これは，顔の下側の動きも考慮するためであり，実際，唇の動きだけでなく，あごや顔の下側の筋肉の動きも音声と関係していることが報告されている．しかし，ひげやしわ，肌の色といった個人差に大きく影響を受けてしまうデメリットもある．

●唇・音声特徴

《唇特徴量》

唇特徴は主に，輪郭ベースの特徴量とアピアランスベースの特徴量に分けられる．前者は，唇領域の横幅，高さなどを用いるものや，唇の輪郭モデルを構築し，唇の輪郭にフィットするモデルのパラメータを特徴量とするものなどがある．後者は，唇領域の画素値に対してPCAや，LDA（Linear

図-3　AAMによる唇領域抽出例

Discriminant Analysis：線形判別分析），DCT（Discrete Cosine Transform：離散コサイン変換）を適用したものなどがある．このように，唇特徴量にはさまざまなものが用いられており，輪郭ベースの特徴量と，アピアランスベースの特徴量を組み合わせて用いている研究も多い．

《任意の顔方位を許す発話》

従来，発話者は顔を正面に固定して発話していたが，近年，任意の顔方位を許して発話認識する研究が行われている．これは，発話者が発声する際に，必ずしも正面を向いて発声するとは限らないからである（たとえばカーナビゲーションでは，ユーザはカーナビを見たり，運転方向を見たりする）．

任意の顔方位に対するアプローチとしては，方位ごとにモデルを用意しておき，発話者の方位を推定した後，その方位に最も近いモデルを選択して，発話内容を認識する方法が提案されている．また，横顔から正面顔への変換行列を学習しておき，横顔からの発話に対して正面顔に変換して認識する方法などが提案されている．

《音声特徴量》

音声特徴量は，MFCC（Mel-Frequency Cepstrum Coefficient）が広く用いられている．MFCCは，基本周

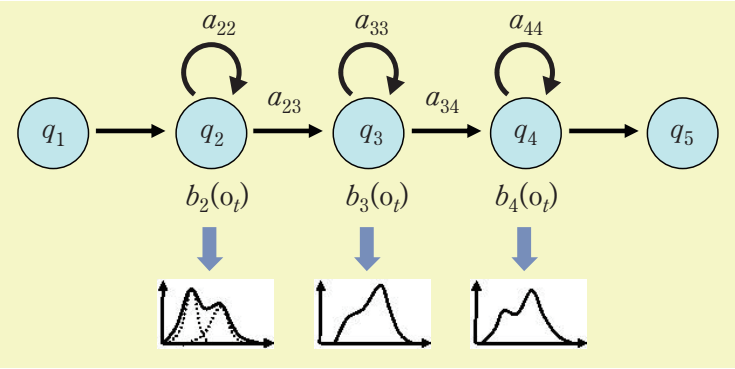

図-4　HMM の例

波数とその調波成分を取り除き，音韻情報を含んでいるスペクトル包絡を抽出したもので，発話変動にロバストな特徴量である．MFCC は，音声信号をメル尺度☆2の周波数軸上でフィルタバンク分析☆3を行い，その出力値の対数を離散コサイン変換することで得られる．

《動的特徴量》

多くの場合，発話認識を行う際には，音素を認識の基本単位としている．上述した MFCC のような特徴量は，数十 ms の区間（フレーム）を定常とみなした上で得られる静的な特徴量である．しかし，音素は，前後に存在する音素の影響を受けて変化するため，静的な特徴に加え，時間とともに変化する動的な特徴を加えることで，認識精度が大きく向上することが知られている．動的特徴としては，静的特徴の時系列に対する回帰係数が広く用いられている．このように，音声特徴量は，静的特徴と動的特徴を結合したベクトルの時系列として表現される．

●発話認識技術

《HMM の基本概念》

音声パターンを見ると，周波数成分が時間方向に変化している．音声パターンのこの時間的・周波数的な変化を，確率的にモデル化したものが HMM（Hidden Markov Model：隠れマルコフモデル）である．典型的な HMMを図-4 に示す．図-4 の HMM には，3 つの状態 q_2, q_3, q_4 があり，それぞれの状態には，時刻 t において音声の特徴ベクトル o_t を生成するための出力確率分布 $b_2(o_t)$，$b_3(o_t)$，$b_4(o_t)$ が存在している．ある状態から次の状態への遷移は，状態遷移確率 a_{ij} に基づいている．これより，HMM は状態を遷移しながら，その状態の出力確率分布に従って，音声の特徴ベクトル o_t を生成していく確率モデルとなっていることが分かる．

この HMM を音声認識に用いるには，たとえば，複数の単語 HMM に対して，入力された音声特徴ベクトル（時系列）の確率を求め，これが最大となる単語 HMM を選択することで実現できる．また，各状態の出力確率分布や状態遷移確率は，EM アルゴリズムに基づく Baum-Welch 法で推定することができる．音声のモデル化においては，図-4 に示すように，左方向への遷移がない Left-to-Right 型を用いるのが一般的である．

《HMM の基本構成》

各状態に設定されている出力確率分布としては，複雑な分布形状を表現するために，複数の正規分布の重み付き和である GMM（Gaussian Mixture Model：混合正規分布）を用いることが多い．このとき，GMM の m 番目の正規分布に関する平均ベクトル μ_m と分散共分散行列 Σ_m，混合分布の重み w_m が，HMM のパラメータとなる．

HMM のパラメータを λ とするとき，状態の遷移を表す状態系列 $Q=\{q_1, q_2, \cdots q_t\}$ に対して，出力ベクトル系列 $O=\{o_1, o_2, \cdots o_t\}$ が出力される確率は

$$P(O, Q \mid \lambda) = \prod_{t=1}^{T} a_{ij} b_j(o_t) \qquad (1)$$

となる．一方，出力ベクトル系列 O が出力される確率は，すべての状態遷移の組合せを足し合わせる

☆2 人間の音高の聴覚特性に対する感覚尺度．
☆3 周波数軸上に配置された複数のフィルタ群の出力に基づいて行う分析のこと．

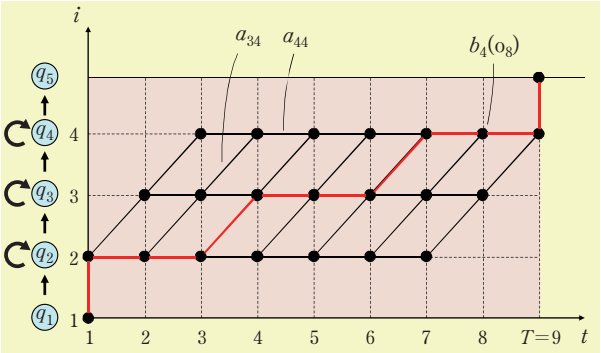

図-5　HMM の状態遷移の例

ことで，

$$P(O \mid \lambda) = \sum_{Q} P(O,\ Q \mid \lambda) \qquad (2)$$

として計算することができる．式(1)は，図-5 において，左下端のノードから右上端のノードへ至る 1 本のパスを考え，このパス上の確率を掛け合わせたものであり，各パスは，それぞれ 1 つの状態系列 Q に対応している．式(2)は，すべてのパスに対応する確率を加え合わせたものであるが，パスを特定する目的で，式(2)の Q による総和の代わりに，$P(O, Q \mid \lambda)$ を最大にする Q を求める Viterbi アルゴリズムが広く用いられている．

《認識単位》

発話認識を行う場合，音素 HMM を連結して単語 HMM を構成し，これを用いて単語認識や連続音声認識を行うことが多い．すなわち，/a/，/i/，/u/，/e/，/o/ といった音素単位で HMM を学習しておき，発話辞書を用いて音素 HMM を連結することで，任意の単語 HMM を構成する．この結果，任意の発話データに対して単語系列を認識して出力することが可能となる．

唇と音声の統合方法

唇・音声統合認識システムを構築する際に問題となるのは，抽出した音声情報と唇の動き情報(以下，唇の動き情報を画像情報と呼ぶ）をどのように統合するかという点と，周囲の環境に応じてどちらの情報を重視するかという点である．以下では，唇・音声統合認識においてよく用いられている統合方法について紹介する．少し長くなるため，統合方法の特徴を表-1 としてまとめているので，そちらを参照されたい．

●初期統合

初期統合は，音声特徴量と画像特徴量を結合し，音声・画像特徴量として認識を行うもので，特徴量レベルでの統合方法である[1]．一般に，音声と画像のフレームレートは大きく異なるため，初期統合の場合，音声と画像でフレームレートをそろえておく必要がある．このため，画像特徴量の時系列に対してスプライン補間などを施し，アップサンプリングを行うことが多い．

	特徴	長所	課題
初期統合	音声と画像の特徴を 1 つにまとめて認識する	音声情報と画像情報の同期を考慮できる	どちらかの情報が歪んだ場合，著しい劣化を招く
結果統合	音声と画像の特徴を別々に処理し，その結果を統合する	音声情報と画像情報の有効性を環境に応じて考慮できる	音声情報と画像情報の同期を考慮できない
マルチストリーム HMM	音声と画像の 2 つの HMM を同期させながら統合する	音声情報と画像情報の同期を考慮できる	音声情報と画像情報の時間的なずれまで考慮できない
中間統合	音声と画像の 2 つの HMM に対して状態同士の積を求める，また状態同士の遷移を許容する	音声情報と画像情報の同期を考慮しつつ，非同期性も許容できる	計算量が多い

表-1　唇と音声の統合方法の特徴

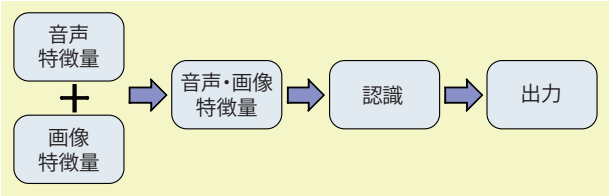

図-6　初期統合

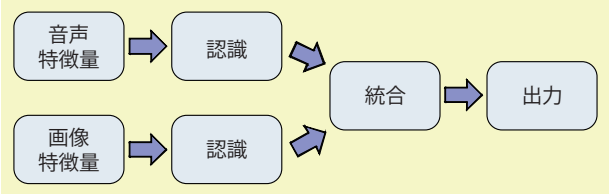

図-7　結果統合

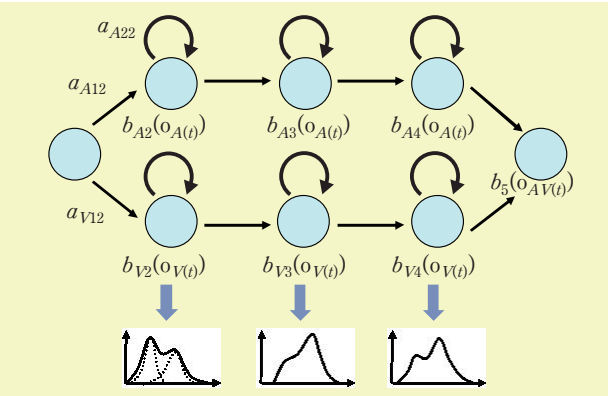

図-8　マルチストリーム HMM の例

図-6 に初期統合の例を示す．初期統合は，最もシンプルな統合方法であるが，音声と画像の特徴量を 1 つのベクトルにするため，特徴量の次元数が大きくなる傾向がある．このため，HMM のパラメータ推定精度が悪くなるといった問題や，音声と画像どちらかの情報が歪んだ場合，著しい劣化を招くといった問題もある．そこで，HMM のパラメータ推定精度を向上させるために，結合した音声・画像特徴量に LDA と MLLT（Maximum Likelihood Linear Transform）を適用して，次元削減した特徴量を用いる方法などが提案されている．

●結果統合

初期統合では，音声と画像の各々の信頼性，すなわち，環境に応じてどちらの情報が有効であるかということが考慮されていない．そこで，音声と画像を別々の過程で処理し，出力された尤度に重み付けを行って統合する．これは結果統合と呼ばれている[2]．図-7 に結果統合の例を示す．結果統合は，音声に重畳する雑音が大きいときは画像の重みを大きくし，音声に重畳する雑音が小さいときや，画像情報に雑音が入ってきたときは音声の重みを大きくするといったように，環境に応じてどちらの情報が有効であるかを考慮することができる．しかし，音声と画像を別々に処理するため，音声情報と画像情報の関連性を音素レベルで有効に利用できていない．また，重みの最適値を決定する必要もある．この問題を解決するために，次節のマルチストリーム HMM が提案されている．

●マルチストリーム HMM

マルチストリーム HMM とは，図-8 に示されるように 2 つのシングルストリーム HMM を融合することにより得られる HMM である[3]．2 つのシングルストリーム HMM とは音声 HMM と画像 HMM であり，マルチストリーム HMM は音声と画像の同期性を前提としている．図-8 のように，特徴量を音声と画像に分割し，各々の信頼性を表現するストリーム重みを用いて，2 つのシングルストリーム HMM を融合する．ただし，音声と画像のフレームレートは同じにしておく必要がある．マルチストリーム HMM の状態 i において，時刻 t の音声・画像特徴量 $o_{AV(t)}$ を観測する確率 $b_i(o_{AV(t)})$ は，

$$b_i(o_{AV(t)}) = b_{Ai}(o_{A(t)})^{\lambda_A} \cdot b_{Vi}(o_{V(t)})^{\lambda_V} \qquad (3)$$

となる．式 (3) において $b_{Ai}(o_{A(t)})$，$b_{Vi}(o_{V(t)})$ はそれぞれ音声特徴ベクトル，画像特徴ベクトルが観測される確率，λ_A，λ_V は音声，画像のストリーム重みである．ストリーム重みを導入することで，シングルストリーム HMM の信頼性を活かすことができる．

マルチストリーム HMM の学習には，HMM のパラメータ（混合分布の重み，平均ベクトル，分散共分散行列，状態遷移確率など）の推定と，適切なストリーム重みの推定が必要である．前者の HMM のパラメータ推定には 2 つの方法がある．1 つは，各シングルストリーム HMM のパラメータを別々に学習し，その結果を式 (3) のように統合する方法であ

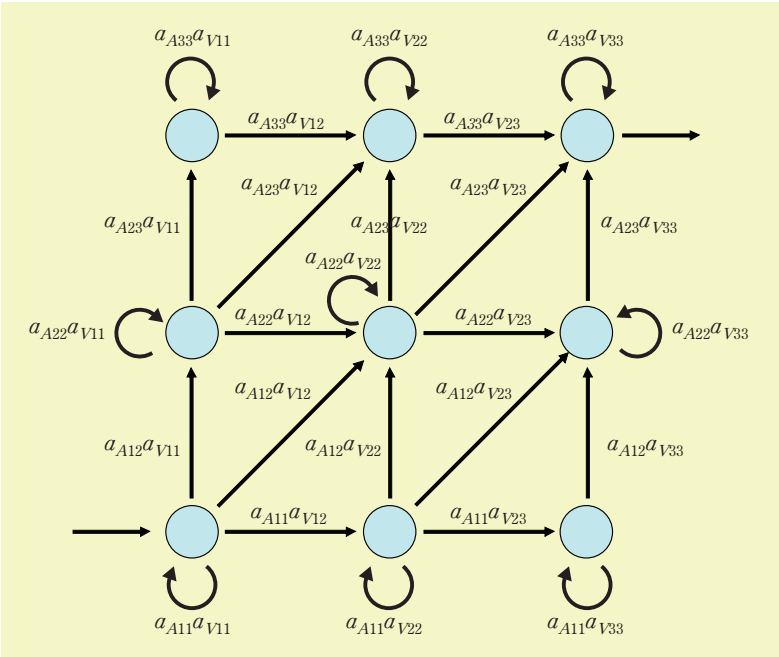

図-9　合成統合の構造

る．この手法は，2つのシングルストリームHMMを非同期で学習するため，式(3)が仮定している状態の同期性に反する．もう1つの方法は，強制的に状態を同期させるために，ストリーム重みを除いたすべてのパラメータを，音声と画像で同時に学習する方法である．この手法は，経験的なストリーム重みの選択を必要とする．しかし，通常，学習時と認識時では環境が異なるため，ストリーム重みを学習によって事前に最適値に設定しておくことはできない．そのため，たとえば最初に適当な重みを設定して音声認識を行い，WER（Word Error Rate：単語誤り率）を最小化するように重みを推定し，もう一度認識するといった方法がとられる．

マルチストリームHMMは，認識時に音声・画像それぞれの雑音状況に応じて重みを変化させることで，認識性能を改善することができる．そのため，ストリーム重みの最適化は，マルチストリームHMMにおいて重要な問題であり，適応的に重みを推定する手法が提案されている．

●中間統合

発話の際，唇の動きは通常，音声よりも先行することが知られている．そこで，音声と画像の非同期性を許容するような中間的な統合方法が提案されている．

《合成統合（Product HMM）》

初期統合の場合，HMMの状態遷移確率は，音声と画像の間で共有されており，マルチストリームHMMでは音声と画像の間で強制的に状態を同期させているため，両者とも時間的なずれが考慮されていない．しかし，上述したように，唇の動きは通常，音声に先行することが知られている．そこで，マルチストリームHMMの状態同期の仮定を緩め，音声と画像の非同期性を許容するようにした合成統合（Product HMM）が提案されている[4]．

図-9に合成統合の概略を示す．合成統合では，各状態における出力確率は，音声HMMと画像HMMの各状態における出力確率の積として，次式のように定義される．

$$b_{ij}(o_{AV(t)}) = b_{Ai}(o_{A(t)})^{\lambda_A} \cdot b_{Vj}(o_{V(t)})^{\lambda_V} \qquad (4)$$

式(4)の $b_{Ai}(o_{A(t)})$，$b_{Vj}(o_{V(t)})$ は，それぞれ状態 i で音声特徴ベクトルが観測される確率，状態 j で画像特徴ベクトルが観測される確率，λ_A，λ_V は音声，画像のストリーム重みである．合成統合において，音声HMMの状態 i から状態 j へ遷移する確率を a_{Aij}，画像HMMの状態 k から状態 l へ遷移する確率を a_{Vkl} とすると，合成統合の遷移確率 $a_{ij,kl}$ は，

$$a_{ij,kl} = a_{Aij} \cdot a_{Vkl} \qquad (5)$$

となる．合成統合のストリーム重み推定については，マルチストリームHMMと同じ手法を用いることができる．

《Coupled HMM》

非同期性を考慮するモデルとして，HMMを改良したCoupled HMM（CHMM）が提案されている[5]．CHMMは，広義にはDynamic Bayesian Networks（DBN）に属している．DBNは隠れ状態を状態変数として表現することでHMMを一般化し，状態同士に複雑な相互依存があることを許容している．

CHMMでは，複数の状態から現在の状態への遷移が存在しており，これにより音声と画像の非同期性を考慮している．しかし，合成統合と同じく，音素の境界では強制的に同期をとる必要がある．また，CHMMは状態が増えるにつれ，計算コストが指数的に増大する．そこで，CHMMの性質を保ったまま，通常のHMMに変換する手法などが提案されている．

《別の視点からのアプローチ》

上述してきた内容は，音声情報と画像情報を「統合する」ものであり，これらのアプローチでは，音声に重畳する雑音が大きい場合，音声情報が有効に利用されず，その認識精度は画像情報に依存してしまう．そこで，画像情報から生成可能な音声信号の候補を複数生成し，実際に観測された雑音重畳音声との間で整合性を評価することで，雑音と音声を分離する方法が提案されている[6]．

この手法ではHybrid Dynamical System（HDS）というモデルを用いて，音声情報と画像情報を複数の線形システムの結合モデルとして表現し，音声と画像間に存在する共起性や時間差といった時間的構造を学習している．

唇・音声統合認識の展望

唇・音声統合認識は，携帯電話やスマートフォンを想定して，横顔の映像から唇の動きを認識する方法や，頬からあごにかかる筋肉の動きによって発生する電気信号（筋電信号）をセンサで読み取り，音の種類を識別する技術など，さまざまな環境を想定して研究が進められている．今後，唇・音声統合認識は，単語認識から連続音声認識へと展開していくと考えられる．

海外では唇・音声統合認識による連続音声認識のための大規模なコーパスが存在する．しかし，日本では音声の大規模なコーパスは存在しても，唇・音声統合認識のための大規模なコーパスは見受けられない．物体認識の研究では，海外のデータベースを用いても問題はないが，言語は国ごとに違うため，より実用化を目指した認識を行うには，日本語のコーパスを用いて研究する必要がある．今後，唇・音声統合認識のための大規模なコーパスが構築されることにより，唇・音声を統合した連続音声認識の進展が期待される．

また，唇・音声統合認識においては，認識手法としてHMMが主として用いられている．音声認識ではHMMの性能限界が報告されており，新たな発話モデルの必要性が認識されつつある．そのため，唇・音声統合認識においても，これらの研究を取り入れていくとともに，音声と画像の同期性を有効に利用できるような改良も必要である．さらに，実環境でより頑健に認識を行うために，照明変動といった環境変化への適応も必要であろう．

本稿では，唇・音声統合認識における基本的な技術を紹介した．唇・音声統合認識がさらに発展し，コミュニケーションのツールとして利用されるよう，今後の展開が期待される．

参考文献

1) Potamianos, G., Neti, C., Luettin, J. and Matthews, I.：Audio-Visual Automatic Speech Recognition：An Overview, in：Issues in Visual and Audio-Visual Speech Processing, Bailly, G., Vatikiotis-Bateson, E. and Perrier, P.(Eds.), MIT Press, 2004(2004).
2) Lee, J.-S. and Park, C. H.：Adaptive Decision Fusion for Audio-Visual Speech Recognition, Speech Recognition, InTech (2008).
3) Morris, A., Hagen, A., Glotin, H. and Bourlard, H.：Multi-stream Adaptive Evidence Combination for Noise Robust ASR, Speech Communication 34, pp.25-40(2001).
4) Kumatani, K., Satoshi, N. and Shikano, K.：An AdaPtive Integration Based on Product HMM for Audio-Visual Speech Recognition, 2001 IEEE International Conference on Multimedia and Expo, pp.1020-1023(2001).
5) Brand, M., Oliver, N. and Pentland, A.：Coupled Hidden Markov Models For Complex Action Recognition. In IEEE International Conference on Computer Vision and Pattern Recognition, pp.994-999(1997).
6) Kawashima, H., Horii, Y. and Matsuyama, T.：Speech Estimation in Non-Stationary Noise Environments Using Timing Structures Between Mouth Movements and Sound Signals, Interspeech 2010, Mon-Ses3-O5：Multi-Modal Signal Processing, pp.442-445(2010).

（平成22年11月2日受付）

■有木康雄（正会員）ariki@kobe-u.ac.jp

神戸大学大学院システム情報学研究科教授．1979年京都大学大学院工学研究科博士課程修了．画像処理，音声情報処理に従事．
http://www.me.cs.scitec.kobe-u.ac.jp/

■駒井祐人 komai@me.cs.scitec.kobe-u.ac.jp

2010年神戸大学工学部情報知能工学科卒業．現在，同大学院システム情報学研究科博士課程前期課程在学中．顔認識の研究に従事．