HMMを用いた音声からの唇動画像合成法

山本英里† 中村 哲† 鹿野清宏†

唇動画像生成の研究はコンピュータエージェントへの応用や聴覚障害者の補助に有用である．音声入力から唇動画像を生成するために，従来ベクトル量子化やニューラルネットによる変換法が用いられてきた．本論文では，音声認識で広く利用されているHMM（隠れマルコフモデル）を音声から動画像への変換に応用する合成法について検討を行う．HMMを用いた合成法の有効性を示すために，以下の3つの比較実験を設定し結果を示す．まず，フレーム単位合成のベクトル量子化法と音素に基づいた変換法であるHMMの合成法を比較する．HMMの合成値はVQの値より誤差で8.7%，差分誤差で32.1%の減少となる．ここでの誤差は音声・唇パラメータの同期データと合成値とのユークリッド距離で定義される．差分誤差は時間差分の誤差で滑らかさの指標となる．次に，唇パラメータの学習方法について，統計的な唇パラメータの学習が合成精度の向上に役立つことを示す．最後に，唇パラメータの合成単位が音素より細かい単位になれば，合成値の誤差も低減することを示す．以上の実験結果から，HMMを用いる合成法は口形が後続音素に依存する音素で大きな誤差を生じることが明らかとなった．そこで後続音素の情報を効果的に採り入れるHMMの合成法を提案し，後続音素情報を用いないHMM合成法と比較を行う．後続音素情報を用いる合成法の実験結果は，元の合成法に対して誤差・差分誤差で，ともに10.5%の改善を示した．

Speech-to-Lip Movement Synthesis by HMM

ELI YAMAMOTO,† SATOSHI NAKAMURA† and KIYOHIRO SHIKANO†

Synthesized lip movement images can compensate lack of auditory information for hearing impaired people, and also contribute to realize a human-like face of computer agents. We propose a novel method to synthesize lip movement from an input speech using HMM. We verify the effectiveness of the synthesis method using HMM by three experiments. In the experiment, error and time difference error between synthesized lip movement images and original ones are utilized for evaluation. The first experiment shows the error of the HMM method is 8.7% shorter than that of the VQ method. Moreover, the time differential error of the HMM method is reduced 32.1% than that of the VQ method. The second experiment shows that statistical training of lip parameter is more efficient for lip movement synthesis than training to select one of lip parameters. The third experiment shows that lip parameters, which are synthesized at a sub phoneme unit, reduce error than the synthesized at a phoneme. Moreover the errors are mostly caused at phoneme /h/ or /Q/, etc. Since those phonemes are characterized by succeeding phoneme, the context-dependent synthesis on the HMM method is proposed. The proposed context-dependent HMM method reduced to 10.5% (10.5%) compared with the original HMM method.

1. はじめに

近年，人間と機械のコミュニケーション技術においてコンピュータエージェントの研究がさかんである．コンピュータエージェントの唇部位の合成はいまだ自然な動きを作るレベルに至っていない．自然な動きの唇画像系列を合成することができれば，エージェントにより人間らしい動きを付加することができる．また，音声を聞くことができない状況でも人間はlip-readingにより視覚的に発話内容を推測することが可能である．聴覚障害者が電話等，音声が唯一の伝達手段である状況下で発話内容を得るとき，自然な動きの唇動画像を出力すれば発話内容の取得に役立つと考えられる．

唇動画像合成に関連する研究として音声の発声機構のモデルパラメータを推定する研究が行われている[1)~3)]．これらの研究は，音声の発声機構を解明することが目的であり，X線，超音波，MRI等比較的低時間分解能の測定装置で収集されたデータを用いて研究されてきた．一方本研究では，Visual Agentの作成や聴覚障害者用の音声の映像化など，応用システムを構築することを目的としており，違和感のない動き

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

を簡便に再現することに重点を置いている．本研究では人間が見て自然な動きであると感じ，かつ，発話内容が推測可能な唇画像系列を合成することが目標である．唇動画像合成には音声からの変換とテキストからの変換が考えられるが，継続時間長など多様な情報を含む音声からの変換に着目する．

音声から唇画像系列を生成するために，すでにベクトル量子化（VQ）やニューラルネットワーク（ANN）を用いる合成法が研究されている[4)~7)]．1フレーム～数フレームのセグメントごとに，逐次音声パラメータを唇パラメータへと変換する点がVQ法やANN法の特徴である．VQやANNで代表される固定フレーム長単位の変換方法に対して，音素系列を基に唇パラメータを合成するHMMを用いた合成法が提案されている[8)~12)]．HMMは音声認識の分野で広く用いられている音素のモデル化手法である．音声認識では，各HMMの尤度を計算し尤度最大のモデルを認識結果とする．この尤度の計算時に最大尤度の状態遷移系列を求めることができる．HMMを用いた合成法は，最尤状態系列に沿って画像パラメータを出力することで唇動画像を合成する．

HMMを用いた合成方法に，音素を単位として唇パラメータを合成する手法がある[9)]．音素を単位とすると，HMM状態と唇パラメータとの対応がないため，音素内の変化が表せない．唇の動きを正確に表現するには音素より細かな単位であるサブ音素（HMM状態）に注目する必要がある．また唇パラメータの学習方法で，口形素（viseme）の知識を利用してデータからパラメータを選択する方法がある．しかし，精度の高い合成値を得るには，統計的に唇パラメータを学習する必要があると考えられる．

本論文では，音声・画像の同期データから統計的に唇パラメータを学習し，サブ音素単位で唇動画像の合成を行う手法（以下HMM法）を提案する．HMM法の有効性を調べるために，音声と画像の同期データを収録して実際に唇パラメータの合成実験を行う．合成動画像の評価は客観的評価尺度として，唇パラメータの合成値と収録値間の誤差と時間差分誤差を用いる．実験では1) HMM法とVQ法，2) 唇パラメータの統計的な学習法と1データから決定される学習法，3) サブ音素を単位とする唇パラメータ合成と音素を単位とする唇パラメータ合成，の3つの比較評価を行い提案法の有効性を明らかにする．

また，HMM法を用いた実験により/h/や/Q/等，後続音素に口形が強く依存する音素で誤差が大きくなることが分かった．本論文では，さらに，後続音素の情報から効果的に唇パラメータを合成する方法（Succeeding Viseme HMM method: SV-HMM法）を提案し，実験でSV-HMM法がHMM法の誤差を低減することを示す．

本論文の構成は以下のとおりである．2章でHMMを用いた音声から動画像への変換方法の具体的なアルゴリズムを説明する．3章では3つの実験に合わせて実際条件を示し，HMM法と他手法による比較実験の結果を述べる．4章ではSV-HMM法の詳細を説明し，実験結果から後続音素の参照効果を検討する．

2. HMMによる唇動画像合成法（HMM法）のアルゴリズム

2.1 HMM法による音声から唇パラメータへの変換

ここでは，音声認識処理で得られるViterbiアライメントを利用して，音声パラメータから唇パラメータへと変換する仕組みについて述べる．音声認識の分野では，入力音声パラメータとHMM（音素に対応する）のマッチングから認識結果を導く．入力音声パラメータの時系列を $O^{sp}=o_1^{sp},o_2^{sp},\cdots,o_t^{sp}$ とする．ある音素HMMを M，その状態遷移系列を $Q=q_1,q_2,\cdots,q_t$ と表記する．1つのHMMは**図1**に示すように，遷移確率 a_{ij} とその初期状態確率そして音声パラメータの出力確率分布 $b_j(o_t^{sp})$ によって特定される．認識結果として音声パラメータ系列 O^{sp} が観測される確率，つまり尤度，$P(O^{sp}|M)$ をモデルごとに計算する．$P(O^{sp}|M)$ が1番高いモデル M が認識結果となる．

$$P(O^{sp}|M)\simeq\max_{Q}\left\{a_{q_0q_1}\prod_{t=1}^{T}a_{q_tq_{t+1}}b_{q_{t+1}}(o_t^{sp})\right\} \tag{1}$$

式(1)のように，尤度最大のHMM状態のみを各時刻で選択し，モデルの最尤遷移経路 Q を算出して近似的に $P(O^{sp}|M)$ を求める手法をViterbiアルゴリズム[13)]といい，最尤遷移経路の状態割当てをViterbiアライメントという．Viterbiアルゴリズムにより最尤遷移経路が求まれば，**図2**に示すように各フレームにつき最尤の音素「HMM」と各時刻でどの状態に遷移したかを示す「状態番号」が決定される．

本論文では最尤遷移経路は1発話を通して計算されるが，現時刻までのデータによって求めることも可能である．しかし発話開始から現時点までの最尤遷移経路では，状態割当ての精度が低くなる．この場合は現時点から数フレーム遅れて処理を行う遅延処理

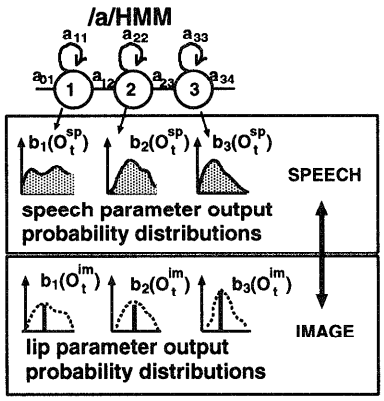


図1 唇画像パラメータの出力確率分布を付け加えたHMMの学習パラメータ
Fig. 1 HMM parameters of speech and image output probability distributions.

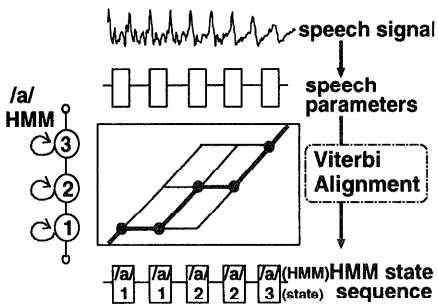


図2 Viterbiアライメント（HMMと状態番号の決定）
Fig. 2 Viterbi alignment (determination of HMM state).

(delayed decision) により，精度を改善することができる．

音声は，HMM状態番号を通して唇パラメータへと変換される．各HMMの各状態 j ごとに，図1に示す唇パラメータ出力確率分布 $b_j(o_t^{im})$ を学習しておけば，各HMM状態で出力確率最大の唇パラメータ値を決定することが可能である．このようにHMM法では，Viterbiアライメントによって各フレームにHMMと状態番号が割り当てられ，状態番号に基づいて唇パラメータが出力される．

2.2 HMM法による唇パラメータ学習アルゴリズム

本論文では，Viterbiアライメントを用いて，データをHMM状態ごとに分類し，唇パラメータの平均値を求める．前節で述べた各HMM状態での唇パラメータ出力確率分布 $b_j(o_t^{im})$ は，HMMを用いてForward-Backwardアルゴリズム[13]から学習できるが，本論文では平均値をViterbiアライメントから求めて，$b_j(o_t^{im})$ の最大確率を与える近似値と見なす．

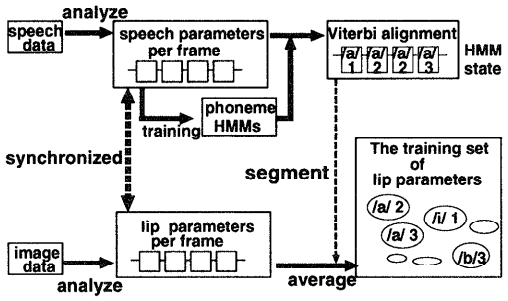


図3 唇パラメータ学習過程
Fig. 3 Training process of lip parameters.

唇パラメータの学習過程を図3に示す．唇パラメータの学習アルゴリズムは以下のとおりである．

学習アルゴリズム

step 1 音声と画像の同期データを用意する．

step 2 音声データと画像データを各フレームで同期をとるよう分析しパラメータ化する．

step 3 学習用音声データのViterbiアライメントをとり，フレームごとに対応する音素ラベル（すなわちHMM）と対応する状態番号を決定する．

step 4 全フレームのうち同じHMMかつ同じ状態番号をとるフレームを選出し，そのフレームすべてに同期した唇パラメータの平均値をとる．

従来提案されているHMM法では，HMM状態ごとの出力パターンを学習せず，凵形のパターンをあてはめている[9]．本論文は上記のアルゴリズムによって，音素HMMすべての状態に対応する唇パラメータをViterbi学習により統計的に求める．

2.3 HMM法による唇パラメータ合成アルゴリズム

次に，テスト音声が入力された後，唇パラメータが合成されるまでの過程を説明する．テスト入力音声でもViterbiアライメントを用い，音素HMMと状態番号の情報から唇パラメータを選択し唇動画像の生成を行う．

唇パラメータの合成過程を図4にブロック図として表す．また唇パラメータの合成アルゴリズムを以下に示す．

合成アルゴリズム

step 1 テスト用の音声データをフレーム単位にパラメータ化する．

step 2 音素HMMを用いて1発話分のViterbiアライメントを取得する．

step 3 HMMと状態番号から，対応する平均値唇パラメータを選択する．

step 4 唇パラメータからフレームごとに3次元の

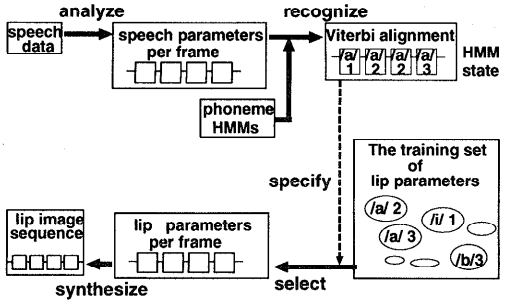


図4 唇パラメータ合成過程
Fig. 4 Synthesis process of lip parameters.

唇画像を合成する.

3. HMM 法による唇動画像合成実験

本章では HMM 法を用いた唇動画像の合成実験について説明する. 3.1 節で画像データと音声データの収録方法や音素 HMM の作成方法等, 実験条件を述べ, 3.2 節にて HMM 法と VQ 法の比較実験を示す. 3.3 節では唇パラメータの学習法の比較を行い, 3.4 節でサブ音素を合成区間の単位とする HMM 法と, 音素を合成単位に用いる唇パラメータ合成法を比較する.

3.1 実験条件

本論文では, 音声の分析周期に合わせて 125 Hz の同期で画像データを収録している. 市販のコンピュータ付属の画像取り込みソフトを使用すると 24〜30 Hz でしか同期をとることができない. 発話画像を測定するにはこの周波数では不十分なため, 3D 位置測定装置を用いて 125 Hz の高レートで音声と同期した画像データを収録する.

3D 位置測定装置は 3 つの CCD カメラを持つ. 収録にあたり被験者は顔に赤外線を発するマーカを貼り, マーカの位置が 3 つの CCD カメラにより 1 平面ずつ走査される. 3 平面の交点から 3 次元位置座標が計測される (**図 5**). 3 次元位置の原点は, 口を閉じた状態で口端点を結ぶラインの中点に設定する.

マーカの張付位置の写真を**図 6** に示す. 顔に貼る各マーカは, 唇外側輪郭の周り 8 点・頬 3 点・顎 1 点の 12 (× 3) channel を用いている. 頭部のマーカは原点同定に用いる. 図 5 のように, 唇上下左右 4 点の channel から画像パラメータとして口の縦幅 X・横幅 Y・奥行き Z の 3 つを作成する.

表 1 にデータの収録条件をまとめる. 音声データは 32 msec 長のハミング窓をかけてフレーム単位に音声パラメータへと変換する. フレームシフトは 8 msec とし, 画像パラメータとの同期を図る. 音声パラメータは [メルケプストラム係数 16 次元]+[差分係数 16 次元]+ [差分 log パワー 1 次元] の計 33 次元を使用する.

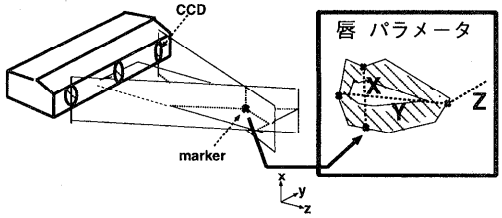


図5 3 平面の交点としてマーカの 3D 位置を決定・唇パラメータの作成
Fig. 5 Sensing markers' 3D positions, and construction lip parameters.

図6 マーカ位置 (唇外側輪郭 8 点, 頬 3 点, 顎 1 点)
Fig. 6 Marker locations (lip outer contour=8 points, cheek=3 points, jaw=1 point) © ATR IIIP.

表 1 音声と画像のデータ収録条件
Table 1 Conditions to record the speech and image data.

画像 sampling rate 画像測定量	125 Hz 各マーカの 3D 位置
画像 marker 数	12 ch
音声 sampling rate	12 kHz
3D 位置測定装置	OPTOTRACK © Northern Digital Inc.
microphone	C−76 © SONY

画像のパラメータは, 前述のとおり口の縦幅 X・横幅 Y・奥行き Z の 3 次元を用いる. この唇パラメータは 3 次元唇画像を再現するための最適パラメータ[14]を参考にした☆.

次にアライメントの取得に使用した HMM の学習条件とテスト単語の認識結果を**表 2** に示す. 音素認識率は, 音素認識により得られる音素列を評価する尺度である. 音素認識率は, Viterbi アライメントの状態割当て精度を近似的に評価する尺度と考えることがで

☆ 文献 14) では唇の内側輪郭の横幅を Y に指定しているが, 3D 位置測定装置では測定不可能なため, 外側輪郭の横の開きを Y に設定した. また 3 次元画像の再現には, さらに上唇の奥行きと下唇の奥行きの 2 パラメータが必要であるが, 簡便のため, 定数とした.

表2 音素HMMの学習条件と認識結果
Table 2 Conditions to train phoneme HMMs and the recognition accuracy.

話者依存性	特定女性話者
モデル数	56 (54音素モデル＋発話前・発話後無声モデル)
状態の接続形態	Left-to-Right 型
状態数	3
文法制約	CV 制約
学習単語数	ATR 音韻バランス単語を含む326単語
テスト単語数	学習単語以外の100単語
音素認識率 (closed)	93.7%
音素認識率 (open)	74.3%

きる．本論文では音素認識率を以下のように定義する．

音素認識率 =(全音素数 − 消去数 − 代替数 − 湧き出し数)/全音素数

上式で，全音素数は全テスト単語に出現する音素の総数を意味する．消去数はあるべきところに音素が現れなかった数を，逆に存在すべきでない音素が湧き出した数を湧き出し数と表す．代替数は別の音素に誤認識した数である．言語モデルの制約として，発話前と発話後に各々の無音モデルを使用する．音声区間では，54通りの音素HMMに，子音（C）は母音（V）と連続で出現する制約（CV制約）を課す．

唇パラメータの学習には，単語326語を用いる．上記音素HMMに正解ラベルを与えてViterbiアライメントを求め，HMM状態ごとに唇パラメータの平均値をとり学習値とする．

唇パラメータの合成後に唇画像を再構築する方法については，文献14)に基づいて行う．3つのパラメータを基に関数を用いて唇の輪郭を形作り，ワイヤーフレームモデルで表現する．

唇パラメータの合成値を評価する尺度として，合成唇パラメータ X_s, Y_s, Z_s と収録唇パラメータ X_o, Y_o, Z_o 間のユークリッド距離による誤差 E（単位cm）を用いる．また，滑らかさの評価には差分誤差 ΔE を定義する．

$$E = \{(X_s - X_o)^2 + (Y_s - Y_o)^2 + (Z_s - Z_o)^2\}^{\frac{1}{2}} \quad (2)$$

$$\Delta E = \{(\Delta X_s - \Delta X_o)^2 + (\Delta Y_s - \Delta Y_o)^2 + (\Delta Z_s - \Delta Z_o)^2\}^{\frac{1}{2}} \quad (3)$$

3.2 HMM法とVQ法の比較実験

3.2.1 VQ法のアルゴリズム

HMM法の有効性の評価のため，VQ法による唇動画像の合成を試みる．以下にVQ法のアルゴリズムを説明する．VQ法ではフレームごとに音声パラメータを唇パラメータへ変換する．パラメータ変換先を学習するために，まずVQによって音声と画像のパラメータで各々コードブックを作成する．次に学習データの各音声コードワードにつき画像コードワードの対応頻度を求める．対応頻度を重み係数として画像コードワードの期待値をとり変換先に指定する．この期待値のとり方により3通りの合成法を試みる．

具体的な学習・合成アルゴリズムを以下に記す．

VQ法の学習アルゴリズム

step 1 学習データの全フレームのパラメータを用いて，音声と画像各々につき，LBGアルゴリズム[13]により代表点256個のコードブックを作る．

step 2 音声の各コードワード C_k^{sp} につき対応する画像コードワード $C_1^{im} \cdots C_{256}^{im}$ の対応頻度 $w_{k,1} \cdots w_{k,256}$ を求める．

step 3 音声の各コードワードにつき対応頻度を重み係数として期待値をとるこの期待値を各音声コードワード C_k^{sp} の変換先画像コードワード $C_{k'}^{im}$ とする．

(1) $C_k^{sp} \xrightarrow[\text{map}]{} C_{k'}^{im} = C_l^{im}$ ($l : w_{k,l}$ が最大の l)

(5) $C_k^{sp} \xrightarrow[\text{map}]{} C_{k'}^{im} = \dfrac{\sum_{l}^{5} w_{k,l} C_l^{im}}{\sum_{l=1}^{5} w_{k,l}}$ ($l : w_{k,l}$ の大きな順)

(256) $C_k^{sp} \xrightarrow[\text{map}]{} C_{k'}^{im} = \dfrac{\sum_{l=1}^{256} w_{k,l} C_l^{im}}{\sum_{l=1}^{256} w_{k,l}}$

対応頻度のとり方によって上式のように3通りの変換先を作成する．(256)は画像コードワードすべてにつき期待値をとった場合，(5)は対応頻度が上位5位の画像コードワードにつき期待値をとった場合，(1)は対応頻度が1番大きな画像コードワードをマッピング先として指定する場合である．

VQ法の合成アルゴリズム

step 1 テストデータの各フレームで音声パラメータから音声コードワード C_k^{sp} を求める．

step 2 音声コードワード C_k^{sp} の変換先の画像コードワード $C_{k'}^{im}$ を出力する．

step 3 出力画像コードワード $C_{k'}^{im}$ のパラメータ値から各フレームで合成画像を作成する．

表3 HMM法とVQ法の誤差
Table 3 Error distances by the HMM method and the VQ method.

	E cm		ΔE cm	
	closed	open	closed	open
VQ (1)	1.43	1.48	0.87	0.88
VQ (5)	1.15	1.22	0.37	0.38
VQ (256)	1.14	1.15	0.28	0.28
HMM（正答時）	1.05	1.04	0.20	0.18
HMM	1.05	1.05	0.20	0.19

3.2.2 実験結果

表3 に学習単語とテスト単語での1フレームあたりの平均誤差 E と差分誤差 ΔE を示す．HMM法については正答ラベルを与えた場合の誤差 E を音素認識時の参考のために“正答時”として表にあげている．正答時と音素認識時の差はViterbiアライメントが正しく行われるか否かの差である．表3より，HMM法の誤差が3種類のVQ法すべてと比べて小さくなることが分かる．音素認識時のHMM法は，VQ法（256）より誤差（E）で8.7%，差分誤差（ΔE）で32.1%の低減を示す．またVQ法では期待値を256画像コードワードのすべてについて求めた変換方法が1番誤差が小さい．HMM法とVQ法（256）の唇パラメータ合成値を**図7**，**図8**，**図9** に示す．図の横軸はフレーム番号による時間を表す．縦軸は唇パラメータ値である．また点線が収録唇パラメータ値を，実線が合成唇パラメータ値を表す．縦の細線は単語の発話始めと発話終了の時刻を示す．口の縦の開き（X）のパラメータは大きく変動し，口の横の開き（Y）はあまり変化しない．奥行き（Z）は原点を基準とする唇の突き出しの移動量なので負の値をとる．HMM法については，図8が正答時のグラフ，図9が音素認識時のグラフである．図9の斜線部分は音素認識間違いを起こした箇所に相当する．

図7～図9からHMM法での合成唇パラメータはVQ法での合成値に比べて滑らかであるといえる．また唇パラメータの総数は，VQ法がコードワード数と同じ256個に対して，HMM法では56HMM × 3状態=168個で済む．

フレーム単位合成のVQ法では滑らかな合成値が望めないので，平滑化手法として過去のコードにより現在のフレームを拘束する方法[1)]等の報告がなされている．

3.3 HMMによる唇パラメータの学習法の比較

唇パラメータの学習法について，HMM法の説明で示した統計的な学習法とは別に，1つのデータを知識を用いて選択する方法がある[9)]．ここでは両学習法について比較実験を行う．

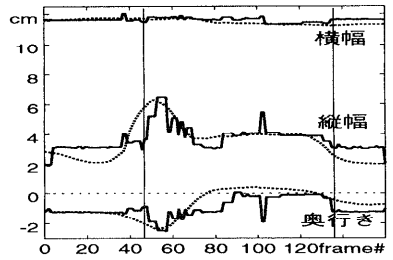


図7 VQ法（256）による合成唇パラメータ/neQchuu/
Fig. 7 Synthetic lip parameters by the VQ method.

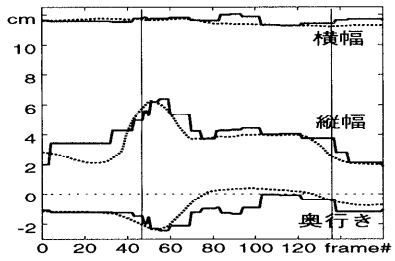


図8 HMM法による合成唇パラメータ（正答時）/neQchuu/
Fig. 8 Synthetic lip parameters by the HMM method with correct phoneme sequence.

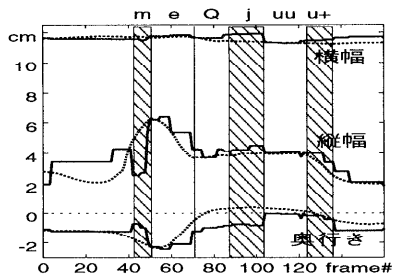


図9 HMM法による合成唇パラメータ（音素認識時）/meQjuuu+/
Fig. 9 Synthetic lip parameters by the HMM method without correct phoneme sequence.

1データから唇パラメータを選択する方法は，以下の手順で行う．まず，HMMを用いて学習データ全部のアライメントをとる．同じ音素に相当するデータの中から，平均的な動きを示すと考えられる1つのデータを目視により選び出す．選択した1データを，HMMの状態番号別に分割し，各HMM状態で平均値を求めて唇パラメータとする．合成は，HMM法の合成アルゴリズムを使用する．

テスト用100単語に対する誤差 E と差分誤差 ΔE を**表4** に示す．ともに音素認識時の結果である．誤差・差分誤差の結果は，データから唇パラメータを統計的に学習する方式の有効性を示している．

表4 唇パラメータ学習法の比較
Table 4 Error distances by HMM training methods.

	E cm	ΔE cm
HMM法（統計的学習）	1.05	0.19
HMM法（1データ学習）	1.33	0.23

表5 唇パラメータ合成区間単位の比較
Table 5 Error distances by lip parameter synthesis unit.

	E(cm)	ΔE(cm)
1音素3唇パラメータ（HMM法）	1.05	0.19
1音素2唇パラメータ	1.08	0.18
1音素1唇パラメータ	1.13	0.15

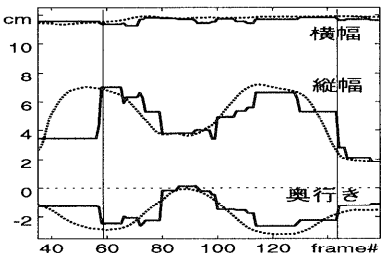


図10 サブ音素単位（3唇パラメータ）の合成/hatsu+ka/
Fig. 10 Synthetic lip parameters based on sub phoneme unit.

3.4 唇パラメータを合成する区間単位の比較

3つ目として，合成の区間単位がサブ音素の場合と音素単位の場合で比較実験を行う．実験は合成区間の細かさで以下の3通りを設定する．

(1) HMMの状態1つにつき1つの唇パラメータを学習・合成（HMM法）
(2) 第1状態で1つ第2，3状態で1つの唇パラメータを学習・合成
(3) HMMで1つの唇パラメータを学習・合成（文献9)）

パラメータの学習方法と合成方法は，HMM法の手法を用いる．HMMはいずれの合成法でも3状態のLeft-to-Right型HMMである．

各合成法でのテスト100単語に対する誤差 E と差分誤差 ΔE を**表5**に示す．また唇パラメータの合成値と収録値の比較の図を，それぞれ**図10**，**図11**，**図12**に示す．

誤差 E は，HMMの状態1つにつき1つの唇パターンを持つ場合に1番良い結果を示す．

比較の図からも，注目の状態数が増えるにつれ収録値に近くなることが分かる．差分誤差は状態数が増えると若干増加しているが，最初にサブ音素単位の合成で厳密な動きを再現しておき，後に平滑化処理を行う方法で滑らかさを補正することが可能と考えられる．

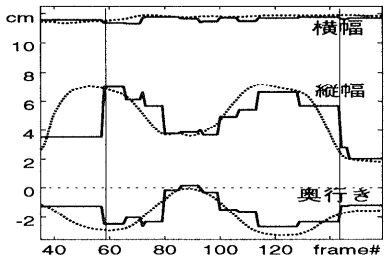


図11 サブ音素単位（2唇パラメータ）の合成/hatsu+ka/
Fig. 11 Synthetic lip parameters based on sub phoneme unit.

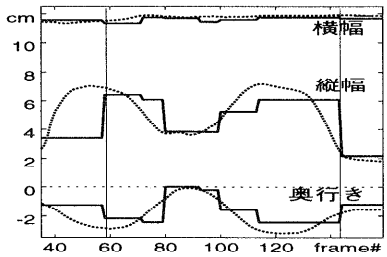


図12 音素単位（1唇パラメータ）の合成/hatsu+ka/
Fig. 12 Synthetic lip parameters based on phoneme unit.

よってサブ音素単位まで正確に合成する方法が音素単位の合成より有効であると考えられる．平滑化処理には，1) 継続時間長を利用する方法[15]や，2)HMMの状態数を増やす方法，3) 複数のHMMの遷移経路を考慮する合成方法，4) 移動平均を用いる方法[9]など，各種方策が考えられる．これら種々の方法については今後検討する予定である．

4. 後続音素依存モデル（SV-HMM法）の導入と実験結果

4.1 後続音素に依存したHMM法

HMM法の実験結果について，誤差の大きな音素を検証したところ，/h/や促音/Q/等の音素で大きな誤差がみられた．**図13**に値の大きな順に7つの音素の誤差をあげる．/h/や/Q/等の音素での唇の動きは，後続音素に依存した動きをとる[16]が，HMM法では後続音素の情報が考慮されないのでこの問題に対処できない．そこでコンテキストを利用し後続音素を参照する方法をHMM法に導入する．

後続音素を参照して後続音素ごとに異なった唇画像を出力するようにHMM法を修正する．特に同じ口形を作る後続音素は1つの口形素にまとめられるので，後続音素を参照する代わりに後続音素の口形素を参照して口形素ごとの唇パラメータを出力することにする．この後続音素の口形素を参照するHMM法を

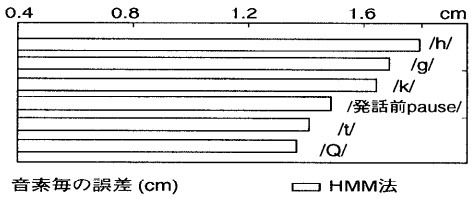


図13 誤差の大きな音素順に並べたフレーム単位の誤差
Fig. 13 Error distances for large error phoneme.

Succeeding Viseme-HMM 法（SV-HMM 法）と呼ぶ．以下に，SV-HMM 法のアルゴリズムを示す．

SV-HMM 法の学習アルゴリズム

step 1 口形素の決定

口形素の種類は，日本語や英語等で各種提案されている[16),17)]．本論文ではデータ量が少ないので，口形素の数は最小限におさえたい．そこで口形素のクラスを3種類に限定する．音素の分類は，前音素の影響が1番大きな各音素 HMM の第1状態にあたる唇パラメータをマージして行う．以下が実験データによる分類結果である．

口形素 1 n b by f m my p py s sh t ts u u+ u- ue ui uu w y z 発話前 pause 発話後 pause

口形素 2 Q ch d g gy hy j k ky n ny o o- oN oo ou r ry

口形素 3 a a- aN aa ai ao e eN ee ei h i i+ iN ii

この結果は，音素の知識から 口形素1が口を閉じた部類，3は口を大きく開けた部類，口形素2は口形素1と口形素3の中間パターンと考えられる．

step 2 学習音声データの Viterbi アライメントをとり，HMM と状態番号を求める．

step 3 各フレームで後続音素の口形素を決定する．

step 4 HMM 状態ごと・後続音素の口形素ごとに唇パラメータの平均値を求める．

SV-HMM 法の合成アルゴリズム

テスト音声データの最尤状態遷移経路を求めて唇パラメータを選択する手順は，HMM 法と同じである．SV-HMM 法では画像パラメータの選択時に「音素 HMM」と「状態番号」と「後続音素」の3つが参照される．選択された唇パラメータは唇画像に変換され，画像系列として出力される．合成のアルゴリズムを以下に示す．

step 1 テストデータの Viterbi アライメントをとり各フレームで HMM と状態番号を決定する．

step 2 各フレームで後続音素の口形素を決定する．

step 3 HMM・状態番号・後続音素の口形素ごとに各フレームに相当する唇パラメータの学習値を出力する．

表6 SV-HMM 法と HMM 法の誤差
Table 6 Error distances by the SV-HMM method and the HMM method.

	E cm		ΔE cm	
	closed	open	closed	open
HMM（正答時）	1.05	1.04	0.20	0.18
HMM	1.05	1.05	0.20	0.19
SV-HMM（正答時）	0.90	0.90	0.18	0.17
SV-HMM	0.91	0.94	0.19	0.17

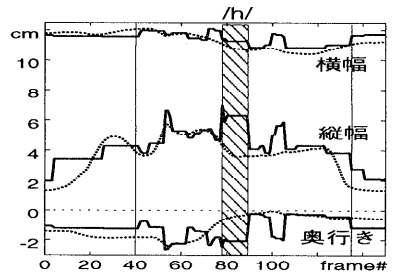


図14 HMM 法による合成唇パラメータ/saki+hodo/
Fig. 14 Synthetic lip parameters by the HMM method.

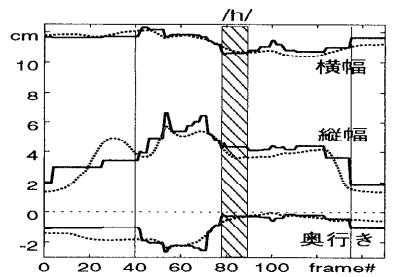


図15 SV-HMM 法による合成唇パラメータ/saki+hodo/
Fig. 15 Synthetic lip parameters by the SV-HMM method.

4.2 SV-HMM 法の実験結果

SV-HMM 法と HMM 法を比較する．実験条件は HMM 法の合成実験と同じである．**表6** に音素認識時での後続音素依存モデルの実験結果を示す．SV-HMM 法は，コンテキストに依存しない HMM 法に比べて誤差が10.5%，差分誤差が10.5%低減する．**図14** と **図15** に/saki+hodo/の正答時の合成唇パラメータを示す．斜線部分が音素/h/に相当する．図14の HMM 法の合成唇パラメータ値は収録唇パラメータ値と大きくかけ離れている．一方後続音素依存モデルの SV-HMM 法を表す図15では/h/での誤差が目立って小さくなっている．

また，この/h/の部分での実際の合成画像を**図16**に示す．左から HMM 法，収録画像，SV-HMM 法に

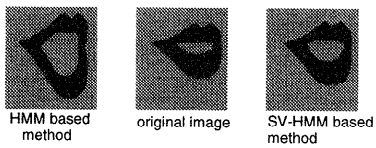


図16 音素/h/での合成唇画像の比較
Fig. 16 Comparison between the synthetic lip images of phoneme /h/.

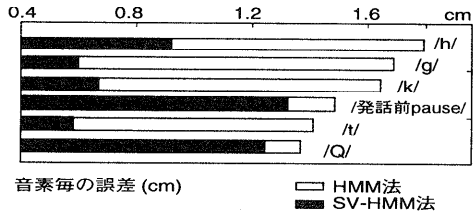


図17 誤差の大きな音素順に並べたフレーム単位の誤差：SV-HMM法の改善
Fig. 17 Error distances for large error phoneme.

おける唇画像である．SV-HMM法の/h/の唇画像が，HMM法の唇画像に比べ収録画像に近い形であることが分かる．また図**17**にSV-HMM法の音素ごとの誤差を示す．図中の音素でSV-HMM法による誤差の改善が見られる．他のすべての音素に対してもSV-HMM法はHMM法の誤差を低減する．以上の結果からコンテキスト依存型の合成方法は音声からの唇動画像合成に有効だと考えられる．

またSV-HMM法と従来法であるVQ法との比較を考えた場合，HMMを用いる合成法はコンテキストを簡単に利用することができるがVQ法は音素認識処理を介さないのでコンテキスト情報を得ることは難しい．しかしVQ法も時間窓を1フレームから前後のフレームを含む複数フレームへと拡張することで，コンテキスト依存に近い効果を出すことが可能である．コンテキスト依存の場合，SV-HMM法では学習する唇パラメータの総数は，56音素 × 3状態 × 口形素数 = 504個となる．一方，VQ法では，時間窓を大きくするとともに，コードブックの学習時間やサイズが増加する．したがって，後続音素依存のHMMを用いた合成法の方が従来法より効率が良いと考えられる．

5. むすび

本論文では音声からの唇画像系列の合成において，HMMを用いて音素単位で音声パラメータから唇パラメータへと変換する手法（HMM法）が，従来手法であるVQを用いたフレーム単位マッピング法（VQ法）に比べて有効であることを客観的評価により示した．HMM法についての実験で，唇パラメータの学習を統計的に行う必要があることを示した．唇パラメータの合成では，音素単位よりサブ音素単位まで注目して合成を行う方法が誤差を低減することを示した．またHMM法において後続音素を参照する合成方法（SV-HMM法）を提案し，参照しない方式（HMM法）に比べて誤差が改善されることを示した．コンテキスト依存型合成法としては，先行・後続両口形素を参照する方法，また調音結合によるコンテキストの影響を合成法に取り入れることが今後の課題である．また今回は発話後に処理を行ったが，実時間処理についても検討する予定である．

本論文では，HMM法の学習過程で決定論的に学習値を求めているが，HMMのForward-Backwardアルゴリズムを用いて非決定論的に推定することも可能である．またHMM法の合成過程では，唇パラメータの合成の質はViterbiアライメントの精度に左右されてしまう．つまりViterbiアライメントが正しく行われないと，間違ったHMM状態の唇パラメータが出力され，大きな誤差を生じる可能性がある．アライメントの精度向上には限界があるので，各フレームで一意にHMM状態を求めるのではなく，すべてのHMMのパスを考慮して確率的に唇パラメータを求める方法を検討していく予定である．

また本論文では客観的誤差のみで合成法を評価したが，視覚的な差異に基づいた主観的評価も行う必要がある．さらに，唇動画像を再構築するために3つの唇パラメータのみ使用しているが，自然な動きを実現するには口の縦幅ではなく上唇・下唇固有の動きを採り入れなければならない．データの解析を通して唇パラメータの最適なセットを決定していく方針である．

謝辞 データの使用を許可していただいたATR人間情報通信研究所の東倉元社長，ならびに，データ収録に協力していただいたBateson博士，ATR知能映像通信研究所の岡田博士に感謝いたします．また唇画像表示ソフトウェアの使用を許していただきましたICPのBenoit博士に感謝いたします．

参考文献

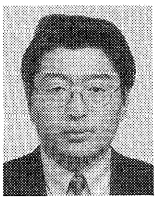
1) Shirai, K., Kobayashi, T. and Yazawa, J.: Estimation of Articulatory parameters by Table Look-Up Method and its Application for Speaker Independent Phoneme Recognition, *ICASSP*, pp.2247–2250 (1986).
2) 白井克彦，誉田雅彰：音声波からの調音パラメータの推定，電子通信学会論文誌，Vol.J61-A, pp.409–416 (1978).

3) Shirai, K. and Kobayashi, T.: Estimation of Articulatory Motion using Neural Networks, *Journal of Phonetics*, Vol.19, pp.379–385 (1991).
4) Morishima, S., Aizawa, K. and Harashima, H.: An Intelligent Facial Image Coding Driven by Speech and Phoneme, *ICASSP 89*, pp.1795–1798 (1989).
5) Morishima, S. and Harashima, H.: A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface, *IEEE Journal on sel. areas in Communications*, Vol.9, No.4, pp.594–600 (1991).
6) Lavagetto, F.: Converting Speech into Lip Movements: A Multimedia Telephone for Hard of Hearing People, *IEEE Trans. on Rehabilitation Engineering*, Vol.3, No.1, pp.90–102 (1995).
7) Curinga, S., Lavagetto, F. and Vignoli, F.: Lip Movement Synthesis using Time Delay Neural Networks, *Proc. EUSIPCO96* (1996).
8) Chen, T. and Rao, R.: Audio-Visual Interaction in Multimedia Communication, *ICASSP 97*, pp.179–182 (1997).
9) Chou, W. and Chen, H.: Speech Recognition for Image Animation and Coding, *ICASSP 95*, pp.2253–2256 (1995).
10) 中村 哲, 山本英里, 永井 論, 鹿野清宏:HMMを用いた音声と唇画像の統合による音声認識と唇画像生成, 情報処理学会研究報告, 97-SLP-15, pp.93–98 (1997).
11) 山本英里, 中村 哲, 鹿野清宏：HMMを用いた音声からの唇画像合成, 第54回情報処理学会全国大会論文集, Vol.2, pp.221–222 (1997).
12) 山本英里, 中村 哲, 鹿野清宏:音声からの唇画像合成におけるコードブックマッピング法とHMM法の比較, 音響学会講演論文集, Vol.I, pp.245–246 (1997).
13) 北 研二, 中村 哲, 永田昌明：音声言語処理, 森北出版, 東京 (1996).
14) Guiard-Marigny, T., Adjoudani, T. and Benoit, C.: A 3-D model of the lips for visual speech synthesis, *Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis* (1994).
15) 益子貴史, 徳田 恵：HMMを用いた唇動画像の生成, 電子情報通信学会技術研究報告, Vol.SP97-6, pp.33–38 (1997).
16) Hiki, S. and Fukuda, Y.: Multiphasic Analysis of the Basic Nature of Speechreading, *Proc. NATO Advanced Study Institute on Speechreading by Man and Machine*, pp.239–246 (1995).
17) Goldshen, A., Garcia, O. and Patajan, E.: Continuous Optical Automatic Speech Recognition by Lipreading, *28th Annual Asilomar Conference on Signals, Systems, and Computers* (1994).

(平成 9 年 7 月 1 日受付)
(平成 10 年 1 月 16 日採録)

山本 英里（学生会員）

昭和 45 年生．平成 5 年奈良教育大学教育学部特別理科物理学専攻卒業．平成 7 年奈良女子大学理学研究科物理学専攻修士課程修了．平成 9 年奈良先端科学技術大学院大学情報科学研究科情報処理学専攻博士前期課程修了．現在，同博士後期課程在学中．マルチモーダルインタフェースの研究に従事．音響学会会員．



中村 哲

昭和 33 年生．昭和 56 年京都工芸繊維大学工芸学部電子工学科卒業．昭和 56～平成 6 年シャープ（株）中央研究所および情報技術研究所に勤務．昭和 61 年～平成元年 ATR 自動翻訳電話研究所に出向．平成 6 年より奈良先端科学技術大学院大学情報科学研究科助教授．平成 8 年 3～8 月 Rutgers University・CAIP Center 客員教授．音声情報処理，主として音声認識の研究に従事．工学博士（京都大学）．平成 4 年日本音響学会粟屋学術奨励賞受賞．IEEE，電子情報通信学会，日本音響学会，人工知能学会各会員

鹿野 清宏（正会員）

昭和 22 年生．昭和 45 年名古屋大学工学部電気学科卒業．昭和 47 年同大学院工学研究科修士課程修了．同年電電公社武蔵野電気通信研究所入所．昭和 59～61 年カーネギーメロン大客員研究員．昭和 61～平成 2 年 ATR 自動翻訳電話研究所音声情報処理研究室長．平成 4 年 NTT ヒューマンインタフェース研究所主席研究員．平成 6 年より奈良先端科学技術大学院大学情報科学研究科教授．音情報処理学講座を担当，主として音声・音情報処理の研究および研究指導に従事．工学博士．昭和 50 年電子情報通信学会米沢賞，平成 3 年 IEEE SP 1990 Senior Award，平成 6 年日本音響学会技術開発賞受賞．IEEE，音響学会各会員．