# ピクセルベースアプローチによる HMMに基づいた唇動画像の生成

酒向 慎司[1]　徳田 恵一[1]　益子 貴史[2]　小林 隆夫[2]　北村 正[1]

[1] 名古屋工業大学 知能情報システム学科
[2] 東京工業大学 大学院総合理工学研究科

[1] 〒466-8555 名古屋市昭和区御器所町
[2] 〒226-8502 横浜市緑区長津田町4259

隠れマルコフモデル(HMM)に基づき，任意の入力テキストから実画像に近い唇動画像を生成するシステムを提案する．我々がこれまでに提案してきたHMMに基づく音声合成法により，高品質なテキスト音声合成システムが実現されているが，これと同一の枠組みを，ピクセルベースの唇画像生成に適用する．音素単位でモデル化された唇動画像HMMを連結し，尤度最大化基準によりHMMの各状態から最適な画像系列を求める．この際，静的特徴量(唇の形状)のみでなく，動的特徴量(唇の動き)を考慮することにより，なめらかに変化する唇動画像を合成することができる．本研究では，新たに作成した日本語連続文章による大規模な唇動画像データベースを用いて，唇動画像合成システムを構築した．任意の入力テキストから合成された唇動画像では，実写画像に近い唇の動きを確認することができた．

隠れマルコフモデル，唇動画像生成，唇画像·音声同期，コンテキストクラスタリング

# Pixel-based Lip Movement Synthesis using HMMs

Shinji Sako[1], Keiichi Tokuda[1], Takashi Masuko[2], Takao Kobayashi[2]
and Tadashi Kitamura[1]

[1]Department of Computer Science, Nagoya Inst. of Tech.
[2]Interdisciplinary Graduate School of Science and Engineering, Tokyo Inst. of Tech.

[1]Gokiso-cho, Shouwa-ku, Nagoya, 466-8555 Japan
[2] 4259, Nagatsuta, Midori-ku, Yokohama, 226-8502 Japan

This paper describes a pixel-based approach for synthesizing lip image sequence from an arbitrarily given text using Hidden Markov Model (HMM). In the training stage, context-dependent lip HMMs are trained and a decisioin tree based clustering technique is applied to them. To synthesize a lip movement, a sentence HMM is constructed by concatenating HMMs corresponding to the transcription for the given text. Then an optimun lip image sequence is obtained from the sentence HMM by using a maximum likeifood criterion. Experimental results show that the synthetic lip image sequence is smooth and realistic.

Hidden Markov model, lip movement synthesis, lip synchronization, context clustering

## 1. まえがき

近年，人間と計算機との間のインタラクションをより豊かにするための研究がさまざまな領域で行なわれてきている．そのなかで，音声などの聴覚情報だけでなく，画像による視覚情報を同時に取り扱う，バイモーダルインタフェースに関する研究が盛んに行なわれている [1][2].

このようなバイモーダルインタフェースは，人間同士のコミュニケーションにより近づいた手段として期待されるが，利用者にとって，モダリティはお互いに独立した情報ではなく，両者は強い相関をもつといわれている．例えば音声と画像の同期がとれていない場合などは，利用者に不自然な感覚を与えるだけでなく，誤った知覚を生じさせ，かえってインタフェースとしての品質を損なうことになる [3].

これまでに，音声と画像の同期に関してさまざまな研究が報告されているが，それらはいくつかのアプローチにまとめることができる [2]．与えられた音声を元に，音声の各セグメントごとに対応する画像を生成する音声駆動型 (speech driven)，またテキストを元にして，各単位ごとに対応する音声と画像を生成するテキスト駆動型 (text driven)，さらにはテキスト·音声駆動 (text-and-speech driven) や画像·音声駆動 (image-and-speech driven) によるハイブリッド型に分類することができる．

音声駆動型では，音声認識やパターンマッチングにより得られたセグメントに対して，あらかじめ学習によって求めておいた画像を割り当てることで，音声と画像の同期を取ることができる．音声から画像へ変換する手法としては，ベクトル量子化 (VQ)[4]，ニューラルネットワーク (NN)[5]，隠れマルコフモデル (HMM)[6][7] 等が用いられる．またテキスト駆動型では入力されたテキストに基づいて音声と画像を合成するもので，一般にトーキングヘッドとして実現されている [8][9]．テキスト駆動は音声駆動よりも一般的なアプローチと言えるが，合成音声の品質がそれほど良くないといった問題が残る．

ところで，HMMは，音声をモデル化する有効な手法として，主に音声認識の分野で広く用いられてきている．我々は，HMMのパラメータから動的特徴量を利用した，なめらかに変化する音声のスペクトル系列を生成するパラメータ生成アルゴリズムと [10]，これに基づいた高品質なテキスト音声合成を提案してきた [11]．また，音声信号だけでなく，動画像のような2次元の時系列信号をモデル化することも可能であり，唇の形状を特徴パラメータとしたモデルベース法による唇動画像合成システムを提案した [12]．唇動画像をモデル化する別の手法として，フレーム毎の画素値を元にパラメータを作成するピクセルベースによる手法もあり，このアプローチでは，データサイズが増加するため，何らかのデータ圧縮を適用する必要があるものの，実画像に近い画像を容易に得ることができる利点がある．

そこで本研究では，実験に必要な日本語連続文章の唇動画像データベースを構築し，HMMに基づく音声合成法と同一の枠組で，ピクセルベースの唇動画像システムを構成した．実験では，与えられたテキストに対応する音素列に従って各唇動画像HMMを連結し，状態継続長分布から各状態の継続長を求め，パラメータ生成アルゴリズムにより唇画像列を生成する．合成動画像からは，実際の発声の動きに近いなめらかな唇画像の変化を確認することができた．

また，音声認識によるセグメンテーションなどによって，実際の音声の音素継続長や状態継続長を求めることができる場合，その認識結果に従ってHMMを連結し，同様の手法でパラメータ生成を行なうことにより，実際の音声に同期した唇動画像を得ることができる．このような音声駆動による唇動画像合成においても良好な結果が得られることを確認した．

以下，2.では動画像生成システムにおけるパラメータ生成アルゴリズム，コンテキストを考慮したモデルの作成，3.では本研究で構成した唇動画像合成システムについて述べ，4.で唇動画像の合成実験とその結果を示し，最後に5.で結論を述べる．

## 2. HMMに基づく動画像合成

HMMに基づく唇動画像生成法のブロック図を，図1に示す．このシステムでは，大きく学習と合成の2つに分けることができる．

学習部では，音声·唇動画像データベースからコンテキスト依存の唇動画像HMMを学習し，2分木に基づくコンテキストクラスタリングによりコンテキスト依存HMMの各状態をクラスタリングして共有化する [13]．さらに，コンテキスト依存HMMの連結学習を行うとともに,連結学習の際のトレリス上で計算される状態継続長から，各状態の状態継続長分布をガウス分布でモデル化し，同様にクラスタリングすることでコンテキスト依存の状態継続長モデルを作成する [14].

合成部では，任意のテキストを入力として，コンテキストに基づいた音素ラベル列に変換し，このラベル列に従って学習部で得られたHMMを連結する．このHMMから状態継続長分布に従って各状態の状態継続長を決定し，得られた状態継続長に基づいて各状態からパラメータ生成アルゴリズムにより唇の画像フレーム列に相当するパラメータ系列を生成する．このようにして得られたパラメータから唇動画像を合成することができる．以下，HMMに基づくパラメータ生成アルゴリズム，決定木に基づくコンテキストクラスタリングについて説明する．

### 2.1 パラメータ生成アルゴリズム

連続出力分布型HMMを$\lambda$で表し，$\lambda$からある状態系列にそって，長さ$T$の出力ベクトル系列$\{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_T\}$を生成することを考える．HMMの各状態は，状態$q$が$d_q$回継続する確率をガウス分布によりモデル化した状態

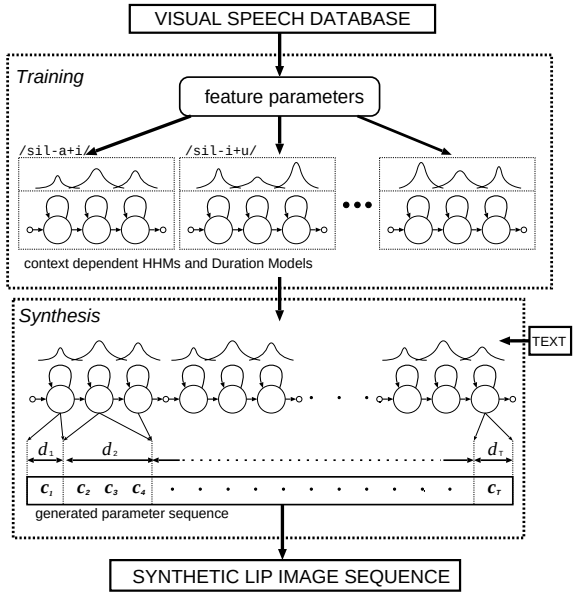



Figure.1　Lip synthesis system

図 1　唇動画像生成システム

継続長分布 $p_d(d_q)$ を持つものとする．また，簡単のため，HMM は単一出力分布型の left-to-right モデルを仮定している．

HMM の状態遷移系列 $\boldsymbol{q} = \{q_1, q_2, \cdots, q_T\}$ にそって出力されるパラメータ系列からなるベクトルを $\boldsymbol{o} = [\boldsymbol{o}_1', \boldsymbol{o}_2', \cdots, \boldsymbol{o}_T']'$ としたとき，与えられた HMM $\lambda$ に対し，出力ベクトル $\boldsymbol{o}$ と状態遷移系列 $\boldsymbol{q}$ の同時生起確率 $P(\boldsymbol{q}, \boldsymbol{o}|\lambda, T)$ の対数は

$$\begin{aligned}&\log P(\boldsymbol{q}, \boldsymbol{o}|\lambda, T)\\ &\quad = \quad a_d \log P(\boldsymbol{q}|\lambda, T) + \log P(\boldsymbol{o}|\boldsymbol{q}, \lambda, T) \end{aligned} \tag{1}$$

と表すことができる．但し，重み $a_d$ は，状態遷移確率 $P(\boldsymbol{q}|\lambda, T)$ と出力確率 $P(\boldsymbol{o}|\boldsymbol{q}, \lambda, T)$ が全体の確率に及ぼす影響を制御するパラメータである．

ここで，HMM は left-to-right であるという仮定から，状態遷移確率 $P(\boldsymbol{q}|\lambda, T)$ は状態継続長分布 $p_q(d_q)$ のみにより決定され，$T$ フレーム中に $K$ 個の状態を遷移するとすると，

$$\begin{aligned}&\log P(\boldsymbol{q}|\boldsymbol{o}, \lambda, T)\\ &\quad = \quad a_d \sum_{k=1}^{K} \log p_{q_k}(d_{q_k}) - \frac{1}{2}\log|\boldsymbol{U}|\\ &\qquad -\frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu})'\boldsymbol{U}^{-1}(\boldsymbol{o}-\boldsymbol{\mu}) - Const \end{aligned} \tag{2}$$

となる．ここで

$$\boldsymbol{\mu} \quad = \quad [\boldsymbol{\mu}_{q_1}', \boldsymbol{\mu}_{q_2}', \cdots, \boldsymbol{\mu}_{q_T}']' \tag{3}$$

$$\boldsymbol{U} \quad = \quad \mathrm{diag}[\boldsymbol{U}_{q_1}, \boldsymbol{U}_{q_2}, \cdots, \boldsymbol{U}_{q_T}] \tag{4}$$

であり，$\boldsymbol{\mu}_{q_t}$ および $\boldsymbol{U}_{q_t}$ は，それぞれ状態 $q_t$ の平均および共分散である．$d_{q_k}$ は，$k$ 番目の状態の継続長を表し，$\sum_{k=1}^{K} d_{q_k} = T$ となる．

$\boldsymbol{q}$ が与えられたとき，$P(\boldsymbol{o}|\boldsymbol{q}, \lambda, T)$ を最大化する $\boldsymbol{o}$ は，$\boldsymbol{o} = \boldsymbol{\mu}$ となる．すなわち，個々のフレームにおける出力は，前後のフレームの出力とは独立に，そのフレームに対応する状態における分布の平均となるため，次の状態へ遷移する部分で不連続が生じる．

これを避けるため，フレーム毎に独立な静的特徴量 $\boldsymbol{c}_t$ と，前後のいくつかのフレームの静的特徴量の線形結合で表される動的特徴量 $\Delta\boldsymbol{c}_t, \Delta^2\boldsymbol{c}_t$ を導入し，$\boldsymbol{o}_t = [\boldsymbol{c}_t', \Delta\boldsymbol{c}_t', \Delta^2\boldsymbol{c}_t']'$ とすることを考える．$\Delta\boldsymbol{c}_t, \Delta^2\boldsymbol{c}_t$ はデルタおよびデルタデルタパラメータと呼ばれ，音声認識で有効なパラメータであることが知られている．これらの特徴量を一般的に

$$\Delta^{(n)}\boldsymbol{c}_t = \sum_{i=-L^{-(n)}}^{L^{+(n)}} w^{(n)}(i)\boldsymbol{c}_{t+i}, \quad n = 0, 1, 2 \tag{5}$$

と表すことにする．ただし，$L^{-(0)} = L^{+(0)} = 0, w^{(0)} = 1$ とする．このとき，静的なパラメータ系列からなるベクトル $\boldsymbol{c} = [\boldsymbol{c}_1', \boldsymbol{c}_2', \cdots, \boldsymbol{c}_T']'$ に関する $P(\boldsymbol{o}|\boldsymbol{q}, \lambda, T)$ の最大化は，

$$\frac{\partial \log P(\boldsymbol{o}|\boldsymbol{q}, \lambda, T)}{\partial \boldsymbol{c}} = \boldsymbol{0}_{TM} \tag{6}$$

として与えられた線形方程式を解くことにより求められる．ただし $\boldsymbol{0}_{TM}$ は，$\boldsymbol{c}_t$ の次元を $M$ として，$T \times M$ 次の零ベクトルとする．さらに，式 (1) の状態遷移確率の重み $a_d$ に十分大きな値を仮定すれば，状態継続長 $\{d_{q_k}\}_{k=1}^{K}$ が，出力確率 $P(\boldsymbol{o}|\boldsymbol{q}, \lambda, T)$ とは独立に状態遷移確率 $P(\boldsymbol{q}|\lambda, T)$ のみにより決定される．このとき，与えられた全体のフレーム長 $T$ に対し，個々の状態の継続長 $\{d_{q_k}\}_{k=1}^{K}$ は，次式によって求められる．

$$d_k \quad = \quad m_k + \rho\sigma_{q_k}^2 \tag{7}$$

$$\rho \quad = \quad \left(T - \sum_{k=1}^{N} m_{q_k}\right) \Big/ \sum_{k=1}^{N} \sigma_{q_k}^2 \tag{8}$$

ここで，$m_{q_k}, \sigma_{q_k}^2$ はそれぞれ状態 $q_k$ の状態継続長分布の，平均と分散である．従って，最適なパラメータ系列は，式 (7)，(8) によって定まる状態系列に従って，式 (6) から得られる線形方程式を解くことにより，一意に求めることができる．

### 2.2　コンテキスト依存モデル

音声や唇の動作に影響を与える要因 (ここではコンテキストと呼ぶ) として，前後の音素環境のほかにもアクセント型, 構文情報などのさまざまな組み合わせが考えられる．これらのコンテキストを考慮したモデルを構築することで，より精度の高いモデルを得ることが期待できる．

しかしコンテキストの種類の増大により，それらの組み合わせは指数的に増加し，モデル当りの学習データが著しく減少することから，モデルパラメータの推定精度

の低下は免れない．また可能なすべてのコンテキストの組み合わせを網羅する学習データを用意することは，現実的に不可能であることから，生成時に学習データ中に存在しないコンテキストの組み合わせが必要となった場合に，対応するモデルを用意できず，パラメータを生成することができなくなるという問題が生じる．

この問題を解決するために，決定木を用いたHMMの状態のコンテキストクラスタリングを導入する．決定木は二分木であり，それぞれの節(node)ごとにコンテキストを2つに分割する質問が用意されている．すべてのコンテキストは根(root)からそれぞれの節の質問に従って木を下り，葉(leaf)のうちのいずれかに達するため，一旦決定木を構築すれば，学習データに存在しないコンテキストの組み合わせにも対応するモデルが一意に決定される．

1. 全てのコンテキストを1つにまとめ，根とする
2. 現在存在するすべての葉に対して，用意されている全ての質問を適用し，分割の前後で最も尤度が増加する葉と質問のセットを選ぶ．尤度変化の最大値が閾値以下であれば終了する．
3. 手順2で選ばれた葉を2つに分割し，新たな葉を2つ作る．古い葉は節となって手順2で選ばれた質問を保持し，新しくつくられた葉に枝を伸ばす．
4. 手順2に戻る

この手順においては，用意する質問が多いほどより多くの分割を考慮することになる．また，尤度の増加が多い質問から順に選択されていくため効率的なクラスタリングが期待できる．

## 3. 唇動画像合成システムの構築

本研究において，図1で示されるようなシステムを構築する上での具体的な要素として，HMMによる唇動画像のモデル化，学習用データベースの作成，特徴パラメータ，唇動画像の合成について説明する．

### 3.1 HMMによる唇動画像のモデル化

発声に伴って変化する唇の視覚的な形状を分類したものに口形素(Viseme)と呼ばれるものがある．これは，聴覚的な分類である音素よりも数が少なく，これを用いて唇動画像をモデル化することも考えられる．

しかし，音素をベースとした音声合成システムの枠組において唇動画像を生成する際には，同じように各音素ごとに対応する唇動画像モデルを作成すると都合がよい．そこで，音素を単位として唇動画像をモデル化した．

### 3.2 音声·唇動画像データベース

一般にHMMの学習には，大量の学習用データが必要となるが，本研究で必要とされるような日本語連続文章による大規模な唇動画像データベースには，現在入手できるもので標準的なものが存在しない．そこで，実験に先立って男性話者1名による音声·唇動画像データベースを構築した．

発話内容はATR日本語連続音韻バランス文503文章，唇画像は，鼻から顎にかけての唇周辺を家庭用ディジタルVTRを用いて話者の正面から撮影したものである．また，同時にDATを用いて音声を収録し，ディジタルVTRの音声と同期を取ることで，高品質な音声データを得ることができる．収録されたデータは，計算機上に取り込み，動画像の口の開閉を基準に1文章づつ切出した．

音声データベースには，発話されている音声のラベル情報が必要となる．しかし，音素レベルのラベリングを行う場合，音声のスペクトルを元に，人手で音素境界を判別することも可能ではあるが，一定の基準で大量の文章をラベリングすることは大変困難である．そこで本研究では，同じ発話内容の他の音声データベースによりトレーニングしたHMMを用いて，今回収録された音声データのセグメンテーションを行い，その結果に基づき音素境界を決定した．

以上の手順により，サンプリングレート48kHzの音声データ，フレームレート29.97fps，720×480画素，RGB各色8bitのカラー動画像による音声·唇動画像データベースを得た．非圧縮時の全データサイズは約91Gbyteとなった．なお，画像フレームがインターレース処理されていることを利用して，スキャンライン方向に分割することで動画像の時間解像度を2倍にすることも可能である．

### 3.3 特徴パラメータ

ピクセルベースにより唇動画像をモデルする際，唇画像を表す特徴パラメータは，PCAなどの分析を施すことで画像の特徴を効率的に圧縮し，特徴パラメータの次元削減を行うことが一般的である．しかし，本研究ではHMMに基づいた音声合成法を唇動画像合成へ適用することの有効性を評価するため，今回は特徴パラメータについては分析を行わず，画素値をそのまま特徴パラメータとして用いることにする．

データベースの画像は輝度や位置のずれなどの変動があり，これらをあらかじめ正規化する．まず，位置の正規化では，位置や形状の変化が比較的少ないと思われる鼻の位置を基準として，1フレーム毎に調整する．調整の方法としては，輝度成分を取り出したモノクロ画像にsobelフィルタリングを施した，連続する2つのフレームの 鼻 を中心とする矩形領域に注目し，フレーム間の相関を最小化する方向に後続するフレームの位置を調節している．そして，正規化された鼻の位置情報をもとに，唇の動き全体をカバーする口元周辺画像(192×144画素)を切り出す．また輝度の正規化については，切り出したフレーム全体の平均値を求めて調整した．

さらに各フレームを4×4ブロックのサブサンプリング

を行ない，48×36 画素の画像とし，ラスタスキャンした 1 フレームの画素全体を，1728 次元の静的特徴量とする．これに Δ パラメータを付加した合計 3456 次元のパラメータベクトルを唇動画像モデルの特徴パラメータとした．

なお，音素単位の唇動画像の学習は音声の音素ラベルに従って行うが，音素区間の短い場合にも十分な学習フレーム数を確保するため，インターレース処理された画像を 2 つに分割することにより 2 倍のフレームレート (約 60fps) の動画像データとして学習に用いた．

### 3.4 唇動画像合成

唇動画像生成では，合成するテキストを音素列に変換して，これに従って音素単位の HMM を連結し，入力文章に対応する 1 つの文 HMM を構成する．学習時に得られた状態継続長モデルにより各音素内の状態継続長を求め，尤度最大化基準に基づいたパラメータ生成アルゴリズムにより，文 HMM の各状態から画像フレーム列に相当するパラメータ系列が出力される．

また，状態継続長が与えられる場合，たとえば，HMM による音声認識などで実際の音声の音素列と状態継続長が求められる場合には，テキストからの合成と同様にして，そのラベル情報を用いて連結した文 HMM からパラメータ生成を行うことにより，実際の音声に同期した唇動画像を合成することができる．このようにして，音声または音声·テキスト駆動による唇動画像合成への応用も可能である．

## 4. 動画像合成実験

作成したデータベースの全 503 文章のうち，前半 450 文章を学習用データとして，残りをテスト文章として用いた．

輝度と位置の正規化が施された 48×36 画素のフレームから求めた動的特徴を含む 3456 次元の特徴パラメータベクトルを用いて，音素ごとに唇動画像 HMM の学習を行う．HMM は 3 状態のスキップなし left-to-right triphone モデルとし，各状態は対角共分散単一ガウス分布をもつものとした．

さらに先行, 当該, 後続音素を考慮したコンテキストクラスタラリングを行い，コンテキスト依存の唇動画像 HMM を作成した．また，HMM の連結学習時に 3 次元ガウス分布の状態継続長モデルを求め，同様にコンテキストクラスタリングを行うことでコンテキスト依存の状態継続長モデルを作成した．

クラスタリングの際，分割を制御する閾値を変化させることで，モデル内のパラメータの共有の度合いを調節したところ，唇動画像および状態継続長モデルの持つ状態の数は表 1のように変化した．なおコンテキストクラスタリングを行う前の総状態数は唇動画像モデルでは 10698，状態継続長モデルは 3566 である．

唇動画像合成では，学習に使用しなかった 53 文章を使用し，各文章に対応した唇動画像を合成した．クラスタリングによる共有の度合いを高くし，状態数を少なくした場合には，唇の動きははっきりしなくなり，全体的に品質が低下した．逆に，状態数を大きくし過ぎた場合には，唇の動きの明瞭さは向上するが，若干の不連続なフレームがみられた．

| threshold | No. of states (%) | |
|---|---|---|
| | lip HMM | duration model |
| 1 | 5591 (52.3) | 2299 (64.5) |
| 10 | 704 (6.6) | 561 (15.7) |
| 100 | 70 (0.7) | 58 (1.6) |

Table.1 Comparison of clustering results with diffrent threshold

表 1 コンテキストクラスタリングによる状態数の変化

図 2に示す 2 つの画像フレーム列は,「都会では，出会う人のほとんどが見知らぬ人である」という文中にある `/t-o-k-a-i-d-e-w-a/` の区間に対応するデータベース中の実画像 (図 2(a)) と，合成された動画像 (図 2(b)，クラスタリング時の閾値はともに 10，数字はモデルの各状態に対応) を 1 フレームずつ並べたものである (フレーム周期は約 16.7msec)．これより，実画像と良く似た唇の動きが，前後の音素環境に応じてなめらかに変化している様子が確認できる[1]．

## 5. むすび

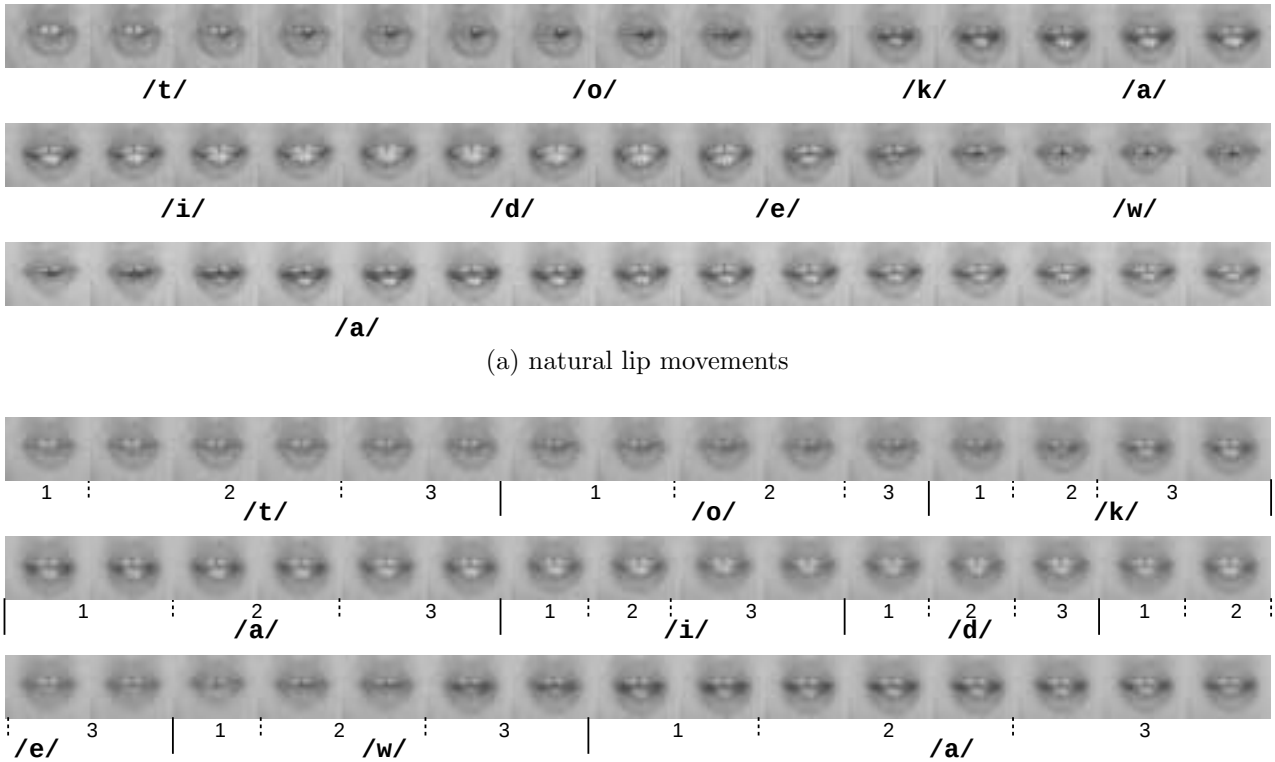
HMM に基づいた音声合成手法と同じ枠組みで，ピクセルベースによる唇動画像生成システムを構築し，実写画像に近い高品質な唇動画像生成が可能であることを確かめた．また，そのために必要な日本語連続文章による音声·唇動画像データベースを構築した．

今回は，唇動画像モデルの特徴パラメータの次元数が画像のサイズに相当するため，モデルの作成や合成時の計算機資源の制約を大きく受けることになった．より現実的な唇動画像合成を行うためには，動画像の特徴をより少ない次元で取り扱う必要がある．今後の課題として，特徴パラメータの次元削減の検討や，音声合成システムを統合した，テキストからの音声·唇動画像生成システムの実現などが挙げられる．

## 謝辞

本研究の一部は，科学研究費補助金奨励研究 (A)(10780226)，萌芽的研究 (11878064) による．また，各種プログラムを提供していただいた名工大博士課

[1] `http://kt-lab.ics.nitech.ac.jp/~sako/lipsynthesis`に，いくつかのサンプル動画像が置かれている




(a) natural lip movements

(b) generated lip movements

Figure.2　Natural and generated lip movements of the sentence “/t-o-k-a-i-d-e-w-a/”

図 2　「都会では」の実際の動画像と，合成した動画像 (1,2,3 は HMM の各状態の継続する区間)

程 吉村貴克氏に感謝致します.

## 文　献

[1] D.G. Stork and M.E. Hennecke, *Speechreading by Humans and Machines*, Springer-Verlag, Berlin, 1996.

[2] F.I. Parke and K.Waters, *Computer Facial Animation*, ch.8 A K Peters, Wellesley, MA, 1996.

[3] H. McGurk and J. MacDonald, “Hearing lips and seeing voices”, Nature 264, pp.746-748, Dec.1976.

[4] S. Morishima, K. Aizawa and H. Harashima, “An inteligent facial image coding driven by speech and phoneme, ” *Proc. ICASSP-89*, pp.1795-1798, 1989.

[5] F. Lavagetto, “Converting speech into lip movements: A multimedia telephone for hard of hearing people,” *IEEE Trans. on Rehabilitation Engineering* , **3**, pp.1-14, 1995.

[6] 中村哲, 山本英理, 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”, 情処研報, 97-SLP-15-7, Feb, 1997.

[7] T. Chen and R.R. Rao, “Audio-vidual interaction in multimedia communication,” *Proc, ICASSP-97, Vol1.I,pp.179-182, Apr. 1997.*

[8] D.R. Hill, A. Pearce, B.Wyvill, “Animating speech: an automated approach using speech synthesised by rule,” *The Visual Computer*, **3**, pp.277-289, 1988.

[9] K. Waters and T.M. Levergood, “DECface: an automatic lip-synchronization algorithm for synthetic faces,” *Technical Report CRL* 93/4, DEC Cambridge Research Laboratory, Cambridge, MA, Sep. 1993.

[10] 徳田 恵一, 益子 貴史, 小林 隆夫, 今井 聖, “動的特徴を用いたHMMからの音声パラメータ生成アルゴリズム,” 日本音響学会誌, vol.53, no.3, pp.192–200, 1997.

[11] 益子 貴史, 徳田 恵一, 小林 隆夫, 今井 聖, “動的特徴を用いたHMMに基づく音声合成,” 信学論, J79-D-II, 12, pp.2184-2190, 1997.

[12] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, “Text-to-audio-visual speech synthesis based on parameter generation from HMM,” *Proc. of EUROSPEECH*, Vol 2, pp.959–962, 1999.

[13] H. J. Nock,M. J. F. Gales and S. J. Yound, “A Comparative Study of Methods for Phonetic Decision-Tree State Clustering,” *Proc. of EUROSPEECH*, pp.111–115, 1997.

[14] 吉村 貴克, 徳田 恵一, 益子 貴史, 小林 隆夫, “HMMに基づく音声合成におけるスペクトル·ピッチ·状態継続長のモデル化,”, 信学技報, S99-59, 1999.