HMMに基づいた視聴覚テキスト音声合成——画像ベースアプローチ

酒向慎司[†] 徳田恵一[†] 益子貴史[††]
小林隆夫[††] 北村正[†]

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，この枠組みを，画像ベースアプローチによる唇画像生成に適用する．これによりテキストから，同期した音声と唇動画像の生成が可能であることを示す．本手法の特徴として，主成分分析によって得られる固有唇を利用して，唇パラメータの次元圧縮を行っている．合成システムでは，連結された唇動画像HMMから尤度最大化基準により最適な唇パラメータ系列を求める．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，連続的に変化する唇パラメータ系列が生成され，それに基づいて，なめらかに変化する唇動画像を合成することができる．

HMM-Based Audio-visual Speech Synthesis —— Pixel-based Approach

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

This paper describes a technique for text-to-audio-visual speech synthesis based on hidden Markov models (HMMs), in which lip image sequences are modeled based on pixel-based approach. To reduce the dimensionality of visual speech feature space, we obtain a set of orthogonal vectors (eigenlips) by principal components analysis (PCA), and use a subset of the PCA coefficients and their dynamic features as visual speech parameters. Auditory and visual speech parameters are modeled by HMMs separately, and lip movements are synchronized with auditory speech by using phoneme boundaries of auditory speech for synthesizing lip image sequences. We confirmed that the generated auditory speech and lip image sequences are realistic and synchronized naturally.

1. はじめに

近年，人間と計算機との間のより豊かなインタラクションの実現を目的として，ヒューマンコンピュータインタフェースに関する研究がさかんに行われている．たとえば，人間と対話できる擬人化されたコンピュータエージェントの実現（たとえば文献1)）は，重要な課題の1つと言える．ところで，人間同士の対話において，特に音声だけでは情報の伝達が不十分な状況において，視覚情報の果たす役割は大きい．たとえば，騒音下や聴覚障害者との対話では，我々は口の動きや身振り，手振りなどの視覚的な情報を利用することで，意志の疎通を行っている．このことから，擬人化エージェントによる対話システムにおいても音声のみの単一のモダリティだけでなく，視覚的なモダリティも同時に扱えることが望ましく，それにより豊かな表現力を持った親しみのあるインタフェースが構築可能であると期待される．

しかし，個々のモダリティは互いに独立した情報ではなく，強い相関を持つといわれている．たとえばモダリティ間の同期がとれていない場合には，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．たとえば，実際の発声と，それとは異なる唇の動きを同時に再生すると，聞き手にはそのどちらでもない音として知覚されてしまうというマガーク効果[2]がその好例である．

このような観点から，本稿では，与えられた任意のテキストから音声だけでなく，音声に同期した唇動画像を生成するテキスト視聴覚（オーディオ・ビジュア

† 名古屋工業大学大学院工学研究科
Department of Computer Science, Nagoya Institute of Technology
†† 東京工業大学大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

ル）音声合成システムを構築することを目的とする．

まず，音声合成のアプローチとして，音声の素片を繋ぎ合わせる波形接続型の手法が一般的に用いられており，これまでに様々なテキスト音声合成システムが提案されている．この手法では，規則的な文章の合成には適しているが，素片の接合部分に歪みが生じやすいなどの問題がある．

顔や唇などの動きを合成するアプローチには，大きく 2 つの手法に分類できる．1 つめのアプローチとして，モデルベースによる手法がある．これは，顔や唇などの形状をモデルパラメータとするもので，表面的な形状だけでなく骨格の動きなども含めた複雑な動きをモデル化することも可能であり，生成されたワイヤフレームのアニメーションから様々な用途に応用できる．3 次元形状をモデル化した顔画像合成なども提案されている[3]．しかし，この手法では，唇内部の舌や歯といった詳細な部分を含めたモデル構築のためには，形状パラメータの抽出にかなりの手間が必要となる．また，合成された唇形状のアニメーションから，実写のようなリアルな動画像を得るためには，テクスチャ画像の貼り付けなど，大がかりなシステムとなる．

もう一方のアプローチとして，フレームごとの画素値を基に特徴パラメータを作成する画像ベースあるいはピクセルベース[4] による手法がある．このアプローチでは，顔画像などを直接モデル化するため，データの取扱いが容易であり，画像にある舌や歯といった詳細な部分も同時にモデル化が可能である．ただし，解像度の高い画像をモデル化する場合には，モデルの特徴次元が増加するため，何らかの次元圧縮を適用する必要があるといえる．

これまでに，我々は音声認識の分野で広く用いられてきている隠れマルコフモデルに基づき，HMM のモデル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[5]，これに基づいたテキスト音声合成システムを提案してきた[6]．この手法の利点として，接続歪みが生じにくく，合成音声の声質を柔軟に変化させることが可能な点があげられる[7]．また，この音声合成手法と同様の枠組みで，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てきた[8]．

本研究では，画像ベースアプローチを採用し，HMM に基づく音声合成の枠組を唇動画像合成に適用する[9]．唇動画像のモデル化には，特徴量として主成分分析を用いて抽出した唇パラメータを用いることで，唇動画像のモデル化が低次元で実現されている．本手法の特徴として，動的特徴を考慮したパラメータ生成によって，滑らかな唇動画像を生成することができる．さらに，音声合成システムと組み合わせ，視聴覚テキスト音声合成システムを構築する．本手法では，音声と唇動画像モデルを音素単位でモデル化し，両者の同期を音素単位で行うことで，任意のテキストから同期した音声と唇動画像の生成が可能である．

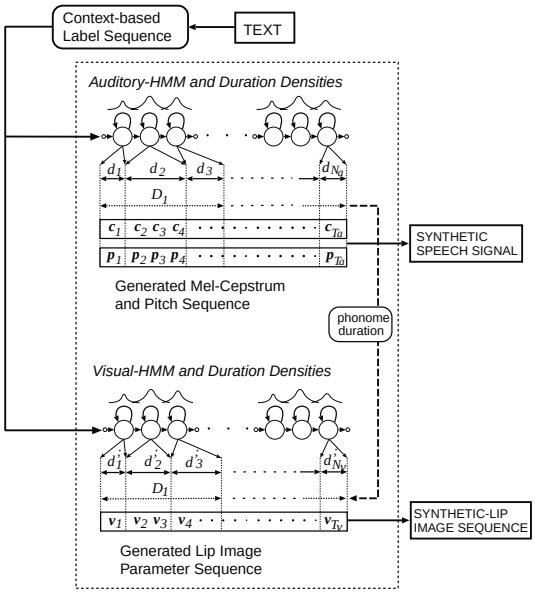


図 1 音声・唇動画像合成システム
Fig. 1 Text-to-audio-visual-speech synthesis system.

以下，本稿は次のように構成されている．2 章では合成システムについて述べ，3 章でデータベースの構築，4 章で音声と唇動画像の合成実験とその評価を行い，最後にまとめる．

2. 視聴覚テキスト 音声合成システム

本稿で提案する，HMM に基づいた音声と唇動画像生成システムのブロック図を，**図 1** に示す．このシステムでは，音声合成部と唇動画像生成部の 2 つから構成されている．音声合成用と唇動画像合成用のモデルは音素単位で個別にモデル化され，それぞれに状態継続長モデルを持っている．

まず，合成したい任意の文章を入力し，ラベル列に変換する．音声合成部では，音声 HMM を連結し入力文章に対応する 1 つの文 HMM とする．そして，状態継続長分布に基づいて音声 HMM の各状態の継続長を決定し，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 の各状態から最適な音声パラメータの系列を生成する．また，唇動画像生成でも同様にして，連結された唇動画像 HMM から唇

パラメータ系列を生成する．生成された音声パラメータと唇パラメータからそれぞれ音声と唇動画像が合成される．

本研究では，音声と唇動画像を別々にモデル化しているが，合成時に両者の音素区間を共有することにより，音声と唇動画像の同期を音素単位で実現している．音声合成時に計算される各音素ごとの状態継続区間の和を音素区間として，唇動画像の合成には，その音素区間に従って唇動画像の状態継続長モデルを用いて各状態継続長を決定する．

2.1 HMM に基づいた音声合成システム

音声合成用のモデルとして，スペクトルパラメータ，基本周波数，および継続長を HMM によって音素単位でモデル化する．音声のスペクトルパラメータにはメルケプストラムを用い，連続 HMM によってモデル化する．また，基本周波数は有声区間では連続値をとり，無声区間では値を持たない可変次元の時間系列信号であるため，通常の連続 HMM や離散 HMM で直接モデル化することはできない．そこで，可変次元に対応した多空間上の確率分布に基づく HMM（Multi-Space probability distribution HMM; MSD-HMM）[10] を用いて，有声音を 1 次元空間，無声音を 0 次元空間のガウス分布でモデル化する．状態継続長モデルは，HMM の各モデルの状態継続長を多次元ガウス分布でモデル化し，状態継続長分布は HMM の連結学習時に作られるトレリス上で求める[11]．

音声の特徴パラメータは，様々なコンテキスト要因によって変動するため，これらを考慮したモデル化を行うことで，より精密な音声モデルを構築することができる．しかし，すべてのコンテキストを網羅した学習データを用意することは現実的ではない．そこで決定木に基づいたコンテキストクラスタリング[12] によってコンテキスト依存 HMM の状態を共有化する．

合成部では，合成したいテキストを入力として，コンテキストベースのラベル列へ変換し，コンテキストに対応するモデルをそれぞれ連結することで，入力されたテキストに対応する文 HMM を構成する．そして，状態継続長分布によって文 HMM の各状態継続長を決定し，パラメータ生成アルゴリズムによって各状態からスペクトルと基本周波数のパラメータ系列を生成する．生成されたこれらの系列を，MLSA フィルタ[13] で励振することにより音声が合成される．

2.2 パラメータ生成アルゴリズム

連続出力分布型 HMM λ と状態遷移系列 $\boldsymbol{Q} = \{q_1, q_2, \cdots, q_T\}$ が与えられるとき，$P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)$ を最大にする音声または画像パラメータ系列 $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_T\}$ を求めたい．ただし 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．

状態遷移系列 $\boldsymbol{Q}$ が状態継続長分布から決定される場合，$P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)$ を最大化するパラメータ系列 $\boldsymbol{O}$ はモデルの平均ベクトル系列と等しくなることは明らかであり，出力されるパラメータ系列は状態ごとに独立して決定されるため，各状態の境界において不連続が生じてしまう．

この問題を解決するため，静的特徴量と動的特徴量から構成される特徴パラメータ $\boldsymbol{o}_t = [\boldsymbol{c}_t', \Delta\boldsymbol{c}_t', \Delta^2\boldsymbol{c}_t']'$ を導入する．なお動的特徴量はデルタパラメータとも呼ばれ，音声認識で有効な特徴量であることが知られている．デルタパラメータは以下のように前後に隣接する静的特徴量 $\boldsymbol{c}_t$ の線形結合によって表される．

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

このような制約の下で，静的なパラメータベクトル $\boldsymbol{c}_t$ からなる系列 $\boldsymbol{C}=\{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_T\}$ は，線形方程式 $\partial \log P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)/\partial \boldsymbol{C} = \boldsymbol{0}_{TM}$ によって与えられ，これは文献 6) に提案されている高速アルゴリズムによって逐次的に計算することができる．

このように動的特徴量を考慮することで，生成されるパラメータベクトルの系列は，前後のフレームの統計量に基づいて生成されるため，連続したパラメータ系列を得ることができる．

2.3 唇動画像のモデル化と合成

画像ベースアプローチでは，解像度の高いデータでモデル化を行うほど，画像の特徴空間の次元が増大するため，効率的な次元の圧縮が不可欠である．情報圧縮の手法には，目的や用途に応じて様々なものがあるが，汎用的な画像圧縮法として，2 次元離散コサイン変換（DCT）やウェーブレット変換などが知られている．また，画像の特徴を低次元で表現する手法として，顔画像認識の分野でよく知られている，固有顔（Eigenface）手法[14] がある．

この手法では，様々な顔画像の主成分分析（Principal Component Analysis，PCA）により固有顔ベクトルを定め，これらの線形結合によって画像を近似した際の，各固有顔ベクトルの係数を特徴パラメータとして利用するものである．ここで，固有ベクトルのうち，固有値の大きいものに画像の特徴の大半が集中している性質を利用することで，画像の持つ情報を効率的に

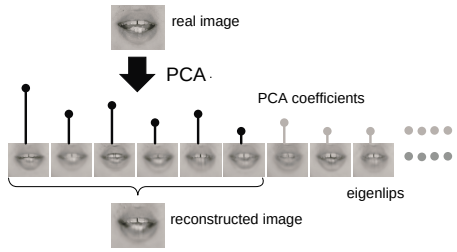


図 2 固有唇による画像の近似
Fig. 2 Lip image approximation using eigenlips.

圧縮できることが示されている．

そこで本研究では，この手法を唇動画像のモデル化に適用し，唇画像から特徴パラメータの抽出法について考える．

1 枚の唇画像の画素値を並べた M 次元ベクトルを $\boldsymbol{x} = [x_1, x_2, \cdots, x_M]$ とし，$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]$ を N 枚の画像フレームからなる画像集合全体の行列とする．ここで N 枚の平均唇画像 $\overline{\boldsymbol{x}}$ と $\boldsymbol{X}$ の各要素との差分を $\widehat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \cdots, \hat{\boldsymbol{x}}_N]$ と表すことにする．

このとき，$\widehat{\boldsymbol{X}}$ の各要素は，唇画像中の変化量の大きい部位が強調されたものと考えることができる．これを様々なパターンからなる多変量データとして，主成分分析を用いて固有ベクトルの集合 $\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_N]$ を求める．これら固有ベクトルは唇画像を構成するための様々な特徴を表していると考えることができ，これらは固有唇（Eigenlip）[15] と呼ばれている．

これらの固有唇を用いて，唇画像 $\boldsymbol{x}$ は $\boldsymbol{x}' = E\boldsymbol{y} + \overline{\boldsymbol{x}}$ のように固有唇の線形結合と平均画像の和として近似することができ，各固有唇にかかる重みベクトル $\boldsymbol{y} = [y_1, y_2, \cdots, y_N]$ は，元の唇画像 $\boldsymbol{x}$ を表現するための各固有唇に対する貢献度であることから，元の唇画像の持つ特徴量として考えることができる．さらに，一部の固有唇のみで近似を行うことで，特徴量の次元を圧縮することができる（**図 2**）．

3. マルチモーダルデータベースの構築

HMM のように統計的手法を用いたモデル構築には大規模なデータベースが不可欠となる．音声研究では様々な研究用データベースが整備されているが，視聴覚研究に関しては各所で整備が進められている段階であり[16]，視聴覚音声合成に適したマルチモーダルデータベースはまだ存在していない．そこで，本研究では日本語連続文章による音声と唇動画像によるマルチモーダルデータベース[17] の整備を行い，実験に使用する．

データーベースの収録は，顔下部（鼻から顎まで）の顔画像を正面から家庭用ディジタルビデオカメラで撮影し，並行して DAT デッキのマイクから音声を録音した．発話する文章は，ATR 日本語音声データベースの音韻バランス文 503 文章を採用し，男性話者 1 名のデータを収録した．

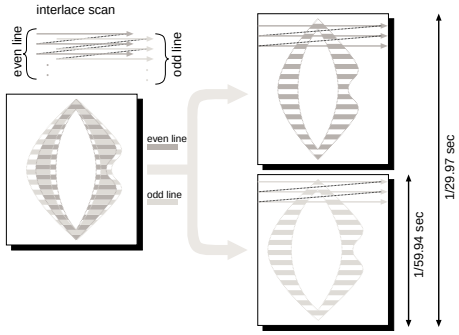


図 3 インタレース画像の分割
Fig. 3 Devide into two frames from interlaced image.

収録されたデータはディジタルデータとして計算機上に取り込んで編集作業を行い，動画像部が 480×720 画素の 24 bit RGB カラー画像，29.97 フレーム/秒，非圧縮で 90G バイト（総フレーム数は 89,726），音声部がモノラル 16bit PCM，標本化周波数 48 kHz のデータを作成した．また，HMM を用いて音声データのセグメンテーションを行い，音素ラベルを作成した．

3.1 実験用データの前処理

このデータベースでは，インタレース走査された画像であることから，**図 3** のように偶数ライン，奇数ラインからなる 2 つのフィールドに分割することができる．これによって，単位時間あたりのフレーム数が多くなり，音素のような細かい時間変化の動きをモデル化する際に有効である．

これらの画像フレームから，データの次元削減のため RGB 色空間から YUV 色空間へ変換し，2 つの色差成分である UV 信号のサブサンプリングを行った．これは，人間の視覚が色差に対して感度が低いことから，色差信号の解像度を下げても視覚的な品質低下を抑えることができ，効率的な情報量の削減が可能なためである．ここでは，色差信号のみブロックサイズ 2×2 でサブサンプルし，1 画素あたりの情報量を 24 bit RGB の半分に相当する 12 bit に削減した．

また，動画像中の顔の位置は厳密に固定されていないことから，いくらか変動があるため，このままでは精密なモデル化を行うには適していない．そこで，1 文章ごとに先頭画像フレームの鼻孔中心位置を手作業でマークし，後続するフレームの鼻孔中心位置を自動でトラッキングすることで位置の正規化を行った．最終的に，トラッキングによって得られた座標を元に，唇の動

き全体をカバーする唇周辺画像 176×144 画素の唇画像を切り出し，これを実験用の唇動画像データとする．

4. 合成実験と評価

4.1 音声モデルの学習

ATR 音声データベースの音韻バランス文 503 文章中の 450 文章（話者は MHT）を音声 HMM の学習用データとし，可変次元のモデルパラメータに対応した MSD-HMM[10] によって，スペクトルと基本周波数をモデル化する．スペクトルパラメータには 0 次を含んだ 25 次メルケプストラム係数を静的特徴量として，1 次と 2 次の動的特徴量を加えた 75 次元ベクトル，また基本周波数も同様にして動的特徴量を加えた 3 次元ベクトルとし，これらの 2 つのストリームからなる合計 78 次元の特徴ベクトルから音声合成用の HMM の学習を行った．HMM は単混合 5 状態の left-to-right HMM でコンテキスト依存モデルとした．なお，動的特徴量は式 (1) に基づいて以下のように計算した．

$$\Delta \boldsymbol{c}_t = \frac{1}{2}(-\boldsymbol{c}_{t-1} + \boldsymbol{c}_{t+1}) \quad (2)$$

$$\Delta^2 \boldsymbol{c}_t = \frac{1}{4}(\boldsymbol{c}_{t-1} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t+1}) \quad (3)$$

コンテキスト依存 HMM から，さらにコンテキストクラスタリングによって HMM の状態を共有化する．また，状態継続長分布を多次元ガウス分布によってモデル化し，同様にコンテキストクラスタリングを適用した．なお，コンテキスト要因には以下ものを用い[10]，スペクトル，基本周波数，状態継続長モデルをそれぞれ別にクラスタリングを行った．

- 文の長さ
- 当該呼気段落の位置
- { 先行，当該，後続 } 呼気段落の長さ
- 当該アクセント句の位置，前後のポーズの有無
- { 先行，当該，後続 } アクセント句の長さ，アクセント型
- { 先行，当該，後続 } の品詞，活用形，活用型
- 当該音素のアクセント句内でのモーラ位置
- { 先行，当該，後続 } 音素

4.2 唇画像の主成分分析

前処理によって得られた画像データ（179,452 枚）から無作為に選別した 1,000 枚の唇画像を用いて 1,000 個の固有唇を求めた．**図 4** は，各固有唇の持つ統計的な貢献度をグラフ化したものである．横軸は固有唇の累積を表し，縦軸はその累積された固有唇の持つ寄与率である．たとえば先頭の 32 個の固有唇は，固有唇の計算に使用した画像全体の約 86%の統計量を持ってることになり，一部の固有唇に多くの統計量が集まっ

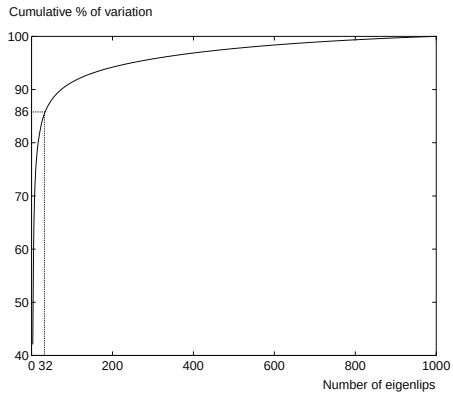


図 4 元画像に対する各固有唇の累積寄与率
Fig. 4 Cumulative variation by eigenlip number.

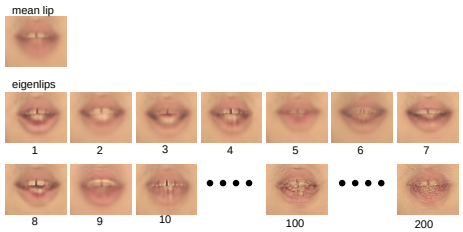


図 5 平均唇と固有唇
Fig. 5 Mean lip image and eigenlips.

ていることが分かる．**図 5** に，固有唇の一部と平均唇画像を示す．ただし固有唇には平均唇画像を加えている．各固有唇には，様々な唇形状が現れていることが確認できる．

4.3 唇動画像 HMM の学習

固有唇によって唇画像を近似する際，固有唇の数が少なすぎる場合には細部の情報が欠落し，全体的に変化の少ないぼけた画像になり，固有唇の数が多いほど細部の近似が可能となる．唇動画像モデルに用いる特徴パラメータの次元数の決定に関して，予備的な実験を行ったところ，次元の増加によって合成された唇画像の品質は向上し，ぼけた画像はより鮮明になっていくものの，増加にともない，合成される唇動画像に，モデルごとの品質の差にばらつきが生じるようになった．合成される唇画像の品質が一定でないことにより，視覚的な違和感を生じやすくなるため，それらの変化の少なかった 32 次元を本実験の条件とした．

予備実験の結果を基に，固有唇を用いてデータベースの唇画像を近似し，各固有ベクトルに対応する 32 次元の係数ベクトルを唇画像の静的特徴量とする．さらに音声と同様にして式 (2)，(3) によって求めた動的特徴量（Δ，Δ^2）を含めた全 96 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

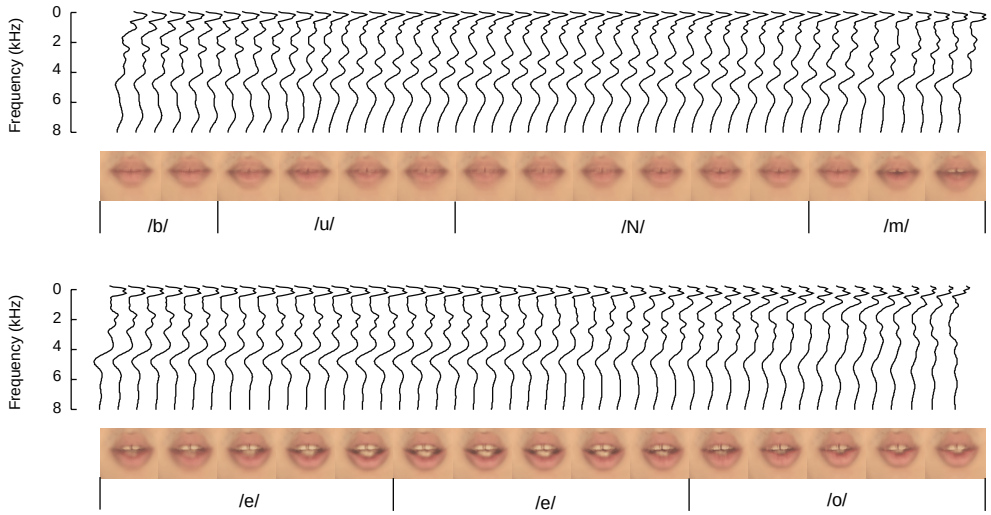


図 7 生成された音声スペクトルと唇画像系列（文明を）
Fig. 7 Synthesized spectra and lip parameters (b-u-N-m-e-e-o).

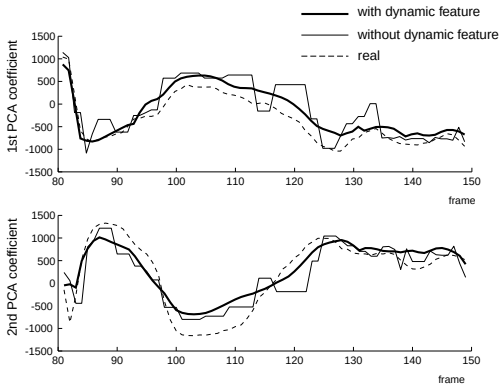


図 6 生成された固有唇パラメータ系列
Fig. 6 Synthesized PCA coefficients vector.

学習を行う．HMM は 3 状態の left-to-right HMM とした．さらに，音声合成用のモデルと同様にして，コンテキスト依存の唇動画像 HMM と状態継続長モデルを作成した．ただし，ここではコンテキストの要因として，前後の音素環境のみを用いている．本来は，唇や舌などの形状変化に関して，音声と同様にして韻律や言語的な変動要因はあると考えられるが，それらに関するデータが未整備であるため，本手法では前後に隣接する音素環境のみとした．

4.4 音声・唇動画像合成

学習データ中にない文章を入力として，音声と唇動画像の合成を行った．**図 6** に，「文明を支える土台が崩れてしまう」という入力テキストから生成された固有唇パラメータ系列のうち，第 1 固有唇と第 2 固有唇に対応するものを示す．グラフの 3 種類の系列は，元画像から求められる唇パラメータと，同一の文章を入力として生成された唇パラメータ（動的特徴あり，なし）である．なお，比較のために実際の音声の継続長を基に生成している．動的特徴を使わない場合には，状態間の遷移において不連続性が見られ，階段状に変化した系列が生成されているのに対し，動的特徴を考慮した場合には，滑らかに変化するパラメータ系列が再現され，元画像から得られる系列によく似ていることが確認できる．また，**図 7** には，合成された音声スペクトル系列と，対応する唇画像系列の一部を示す．この図より，動的特徴量の効果によってなめらかに変化する，スペクトル系列と唇画像列が生成されていることが分かる☆．

まず，動的特徴の効果を確かめるために，試験 1 として動画像の品質について主観評価試験を実施した．試験内容は，53 文章中から無作為に 10 文章を選び，動的特徴なし，動的特徴あり（Δ のみ，$\Delta+\Delta^2$）でそれぞれ生成された動画像を対比較試験により評価させた．なお被験者は 8 名とした．ただし，音声についてはいずれも動的特徴量ありの合成音を使用している．

次に，試験 2 として，音声に唇動画像を付加することによる効果を確かめるために，単音節による明瞭度試験を実施した．単音節には，比較的聞き取りにくいと思われる子音 b，d，g，k，p，t を含むものを選び，音声のみ，音声 + 唇動画像（動的特徴 Δ，Δ^2 あり）の条件でそれぞれ合成した．テストは，一定間隔

☆ 作成された動画像のサンプルがここに置かれている．
http://kt-lab.ics.nitech.ac.jp/~sako/research/

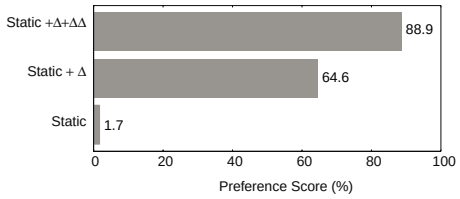


図 8 主観評価試験による動的特徴の効果
Fig. 8 The effect of dynamic features on the subject experiments.

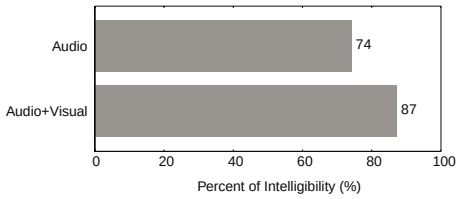


図 9 単音節明瞭度試験による視聴覚音声合成の効果
Fig. 9 The effect of audio-visual speech synthesis on the syllable articulation test.

で再生される単音節を視聴させ，発声された音節を書きとらせる方法とした．1 名あたりのテストサンプルはそれぞれ 25 個をランダムに選択し，合計 8 名に実施した．なお，本手法の合成システムでは，韻律や言語的なコンテキストを考慮しているため，類似単語などで比較する場合には，品詞やアクセントの位置による明瞭性の違いなどを考慮に入れる必要があり，実験条件や評価方法について十分な検討が必要である．そのため，より単純な方式として単音節による評価を実施した．

以上 2 つの試験から得られた結果をそれぞれ**図 8**，**図 9** に示す．まず，図 8 より，音声合成の場合と同様にして[11]，動的特徴量の効果が明確に表れていることが確認できる．動的特徴量を考慮しない場合には，合成されるパラメータ系列の状態間の歪みが大きく，合成される動画像は不連続性の大きいものになってしまうことが評価を大きく下げていると考えられる．次に，図 9 の結果から，音声単独よりも唇動画像を付加することで単音節の正解率が 13%向上していることから，視聴覚音声合成の効果が確認できる．

5. ま と め

HMM に基づいた音声合成手法の枠組みを適用した，画像ベースアプローチによる唇動画像生成システムを提案し，これと音声合成システムと組み合わせることで，任意のテキストから同期のとれた音声と唇動画像の同時生成が可能であることを示した．本手法では，主成分分析を利用した固有唇の導入により，唇画像を固有空間へ写像することで，画像の特徴パラメータの大幅な次元圧縮を可能としている．さらに動的特徴量を用いたパラメータ生成アルゴリズムによって，なめらかに変化する唇パラメータ系列が生成され，連続的に変化する唇画像系列が得られることから，固有唇に基づいた動画像合成への動的特徴の効果を確認した．一方，本手法の問題として，全体的にぼけた画像が合成されてしまう点がある．品質の向上を図るためには，唇画像を近似する際に固有唇をより多く使用するほうが望ましいが，その場合にモデル学習が適切に行われなくなることが分かっている．その対策として唇の形状変化を考慮したコンテキスト要因を導入し，より精密なモデル化を行うことが有効であると考える．

今後の課題として，音声と動画像の関係を正確にモデル化するために両者を単一の枠組みでモデル化することがあげられる．これにはより高いフレーム周期の動画像データの収録や，付随する言語情報などデータベースの整備が必要となる．また，顔全体のモデル化や合成などがあげれらる．

謝辞 本研究で使用したデータベースの収録にあたられた東工大近藤重一氏，また，各種プログラムを提供していただいた名工大博士後期課程吉村貴克氏の両名に感謝いたします．

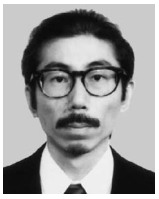
参 考 文 献

1) IPA 独創的情報技術育成事業「擬人化音声対話エージェント基本ソフトウエアの開発．`http://iipl.jaist.ac.jp/IPA/`
2) McGurk, H. and MacDonald, J.: Hearing lips and seeing voices, *Nature*, Vol.264, pp.746–748 (1976).
3) Cohen, M.M., Beskow. J. and Massaro, D.W.: Recent developments in facial animation: An Inside View, *Proc. AVSP*, pp.201–206 (1998).
4) Brooke, N.M. and Scott, S.D.: Two-and Three-dimensional audio-visual speech synthesis, *Proc. AVSP*, pp.213–218 (1998).
5) 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).
6) 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（D-II），Vol.J79-D-II, No.12, pp.2184–2190 (1997).
7) Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T: Speaker Interpolation for HMM-based Speech Synthesis System, *J Acoust. Soc. Jpn.* (*E*), Vol.21, No.4,

pp.199–206 (2000).
8) Tamura, M., Kondo, S., Masuko, T. and Kobayashi, T.: Text-to-audio-visual speech synthesis based on parameter generation from HMM, *Proc. EUROSPEECH*, Vol.2, pp.959–962 (1999).
9) Sako, S., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T.: HMM-based Text-to-audio-visual Speech Synthesis — Image-based Approach, *Proc. ICSLP*, Vol.3, pp.25–28 (2000).
10) Tokuda, K., Masuko, T., Miyazaki, N. and Kobayashi, T.: Hidden Markov models based on multi-space probability distribution for pitch pattern modeling, *Proc. ICASSP*, Vol.1, pp.229–232 (1999).
11) 吉村貴克，徳田恵一，益子貴史，小林隆夫，北村正：HMMに基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化，信学論（D-II），Vol.J83-D-II, No.11, pp.2099–2107 (2000).
12) Nock, H.J., Gales, M.J.F. and Yound, S.J.: A Comparative Study of Methods for Phonetic Decision-Tree State Clustering, *Proc. EUROSPEECH*, pp.111–115 (1997).
13) 今井　聖，住田一男，古市千枝子：音声合成のためのメル対数スペクトル近似（MLSA）フィルタ，信学論（A），Vol.J66-A, No.2, pp.122–129 (1983).
14) Turk, M.A. and Pentland, A.P.: Face recognition using eigennfaces, *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp.586–591 (1991).
15) Brook, N.M. and Scott, S.D.: PCA image coding schemes and visual speech intelligibility, *Proc. Institute of Acoustics*, pp.123–129 (1994).
16) Nakamura, S.: Overview on Recent Activities in Multi-Modal Corpora, *COCOSDA Workshop* (2000).
17) 酒向慎司，近藤重一，徳田恵一，益子貴史，小林隆夫，北村　正：音声と唇動画像によるマルチモーダルデータベースの構築，音響学会講演集，No.3-P-30, pp.223–224 (2001).

(平成 13 年 11 月 19 日受付)
(平成 14 年 4 月 16 日採録)

酒向　慎司（学生会員）
平成 13 年名古屋工業大学工学部知能情報工学科卒業．現在同大学大学院博士後期課程在学中．視聴覚音声合成の研究に従事．日本音響学会会員．



徳田　恵一（正会員）
昭和 59 年名古屋工業大学工学部電子工学科卒業．平成元年東京工業大学大学院博士課程修了．同年東京工業大学電気電子工学科助手．平成 8 年名古屋工業大学知能情報システム学科助教授．工学博士．音声分析，音声合成・符号化，音声認識，ディジタル信号処理，マルチモーダルインタフェースの研究に従事．平成 13 年電気通信普及財団賞，平成 13 年電子情報通信学会論文賞，猪瀬賞各受賞，日本音響学会，人工知能学会，IEEE 各会員，ISCA 各会員．

益子　貴史
平成 5 年東京工業大学工学部情報工学科卒業．平成 7 年同大学大学院博士前期課程修了（知能科学専攻）．同年同大学精密工学研究所助手．現在同大学大学院総合理工学研究科物理情報システム創造専攻助手．音声の分析・合成，音声認識の研究に従事．日本音響学会，IEEE，ISCA 各会員．

小林　隆夫（正会員）
昭和 52 年東京工業大学工学部電気工学科卒業．昭和 57 年同大学大学院博士課程了．同年同大学精密工学研究所助手．同助教授を経て平成 10 年東京工業大学大学院総合理工学研究科物理情報システム創造専攻教授．工学博士．ディジタルフィルタ，音声の分析・合成・符号化・認識，マルチモーダルインタフェースの研究に従事．平成 13 年電気通信普及財団賞，平成 13 年電子情報通信学会論文賞，猪瀬賞各受賞，日本音響学会，IEEE，ISCA 各会員．

北村　　正（正会員）
昭和 48 年名古屋工業大学工学部電子工学科卒業．昭和 53 年東京工業大学大学院博士課程修了．同年東京工業大学精密工学研究所助手．昭和 58 年名古屋工業大学工学部電子工学科講師．昭和 59 年同助教授．平成 7 年名古屋工業大学知能情報システム学科教授．工学博士．音声情報処理，マルチメディア情報処理の研究に従事．日本音響学会，IEEE，ISCA 各会員．