HMM に基づいた音声 · 唇動画像の同時生成 –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　　(名工大, † 東工大)

1. はじめに

我々は，これまでに HMM に基づいた音声合成システムを提案した [1]．本システムでは，動的特徴量を利用したパラメータ生成アルゴリズム [2], [3] により滑らかなスペクトル系列の生成が可能である．また，この音声合成の枠組を，モデルベースアプローチによる唇動画像合成に適応し，テキストから音声と唇のアニメーションを生成するシステムを提案した [5]．

本研究では，これまでに提案した HMM に基づく音声合成システムを画像ベースアプローチによる唇動画像生成に適用する．画素値を元に唇動画像をモデル化することで，任意のテキストからリアルな唇画像系列を合成する．さらに，音声合成システムと組み合わせることでテキストから音声と唇動画像の同時生成が可能となり，同期の取れた音声と唇動画像の生成が可能であることを示す．

2. HMM に基づく音声 · 唇動画像の生成

図 1 に，本研究で提案するシステムの合成部のブロック図を示す．合成部は，音声合成部と唇動画像合成部の 2 つに分けられる．まず，合成したい任意のテキストを音素列に，さらに前後のコンテキストに対応したラベル列に変換し，対応する音声 HMM を連結することで 1 つの文 HMM を構成する．学習時に求められた状態継続長分布モデルから，各音素の状態継続長を決定し，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って音声 HMM，唇動画像 HMM から，それぞれ音声パラメータと唇パラメータを生成する．音声と唇動画像は別々のモデルから生成されるが，音声合成部で生成された音素継続長に従って唇動画像の状態継続長を決定することで，同期のとれた音声と動画像を合成することができる．

2.1 音声のモデル化

メルケプストラムとピッチを同時にモデル化することを考える．ピッチは有声区間では 1 次元空間，無声区間では 0 次元空間という可変次元の観測値を取るため，通常の HMM では取り扱うことができない．そこで，可変次元の多空間上の確率分布に基づいた HMM によりピッチをモデル化する [1]．また，状態継続長は多次元ガウス分布でモデル化する．HMM に状態継続長分布を含めて学習する場合，計算時間が非常にかかるため，各状態の継続長分布を HMM の連結学習時に作られるトレリス上で求める．生成されたメルケプストラムとピッチ系列を，MLSA フィルタによって励振することで音声を合成する [4]．

2.2 唇画像のモデル化

唇画像合成のアプローチとして，唇の位置的な形状を特徴パラメータとして唇をモデル化するモデルベースアプローチ [5] と，画像の画素値を元にしたパラメータで唇をモデル化する画像ベースアプローチ [6] の 2 つがある．

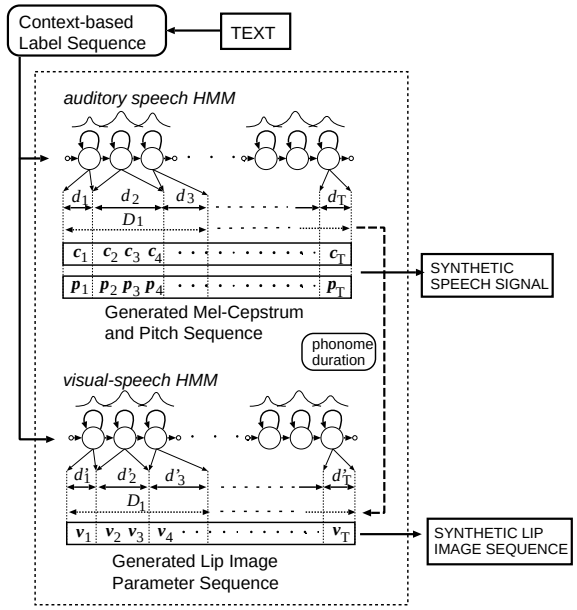

図 1．音声 · 唇動画像生成システム

モデルベースアプローチでは，位置や形状情報をもとに 3 次元的な構造を精密にモデル化することが可能である反面，モデルパラメータとなる特徴点の抽出が困難であるという問題がある．またリアルな画像を合成する場合，唇形状のアニメーションの生成のほかに，コンピュータグラフィクスなどを作成する必要もある．

これに対して，画像ベースアプローチでは特徴パラメータ次元が大きくなる反面，画像を直接利用するため，学習データを比較的容易に作成することができる．また，唇の動きだけでなく歯や唇といった唇の内部を同時にモデル化できるため，生成時には，それらも含めた合成が可能となる．画像の解像度が高くなるほどパラメータ次元が増大するが，本研究では顔画像認識で用いられている固有顔 (eigenface) 手法 [7] と同様，主成分分析に基づいた，固有唇 (eigenlip) 手法により次元圧縮を行っている．生成された固有空間上のパラメータ系列から，元の高次元空間へ写像することで，唇画像系列を合成することができる．

2.3 音声 HMM と唇動画像 HMM の学習

音声合成用の HMM は，ATR 日本語連続音韻バランス文章データベースの男性話者 1 名を用いた．唇動画像ではこのような標準的なデータベースが入手できないため，音声データベースと同一の文章による唇動画像データベースを DVC と DAT レコーダを利用して作成した．収録されたデータは，画像部が 24bit カラーの 720×480 画素，フレームレートが 29.97fps である．音声部は，DV と同期を取った DAT の音声を利用し，16bitPCA，標本化周期は 48kHz とした．さらに音素 HMM によるセグメンテーション

∗ HMM-based audio-visual speech synthesis – image-based approach –.

By Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko†, Takao Kobayashi† and Tadashi Kitamura (Nagoya Institute of Technology, †Tokyo Institute of Technology)

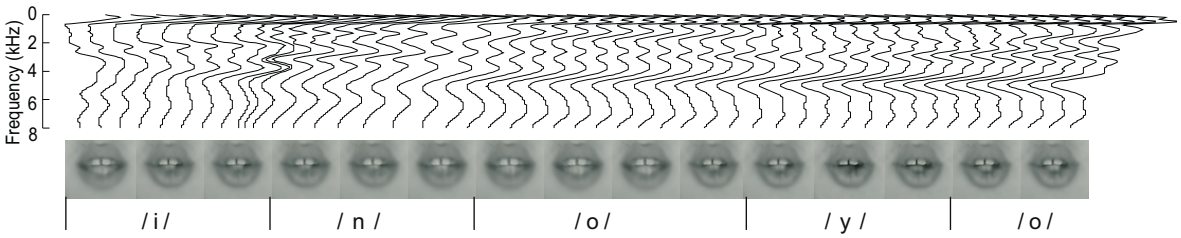

図 2．生成されたスペクトルと唇画像系列の一部 “/i-n-o-y-o/”

の結果に基づいて，半自動的に音素ラベリングを施した．これらのデータベースのうち，それぞれ 450 文章を用いて，音素単位による音声 HMM と唇動画像 HMM を学習した．唇動画像をモデル化する単位として，viseme を用いることも考えられるが，ここでは音声と同一の音素単位を用いることで，容易に合成音声との同期をとることを可能としている．

音声の特徴パラメータはメルケプストラム分析で得られた 25 次元ベクトルに動的特徴量として Δ，Δ^2，またピッチとそれの動的特徴量を含んだ 78 次元 (無声部では 75 次元) ベクトルとして，各音素を 5 状態 left-to-right HMM で学習する．さらに音素または言語に関する質問を適用したコンテキストクラスタリングによってコンテキスト依存の triphone モデルを構成した．また，コンテキスト依存の状態継続長モデルを同時に構成した．

一方，作成した音声 · 唇動画像データベースから，唇周辺を 176×144 画素で切り出した 256 階調のモノクロ画像に位置と輝度の正規化を施し，ランダムに選んだ 1024 枚の唇画像に対して主成分分析を適用し，1024 個の固有唇を求めた．唇動画像の各フレームは，いくつかの固有唇の線形結合で表すことができ，使用する固有唇の数によって次元を圧縮することができる．本研究では，対応する固有値の大きいものから 32 個の固有唇を用い，それらに対応した係数を唇パラメータとして使用する．動的特徴量として Δ, Δ^2 を加え，計 96 次元のベクトルを特徴パラメータとして，3 状態 left-to-right HMM を学習した．さらに音素に関する質問を適用してコンテキスト依存の triphone HMM を構成し，また，状態継続長モデルを構成した．

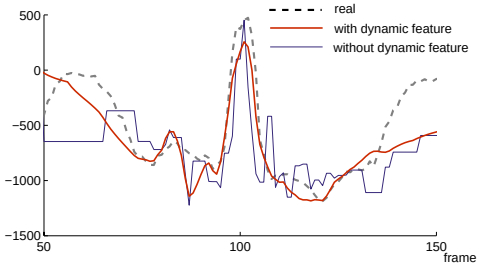

図 3．生成されたパラメータ系列における動的特徴量の効果

3. 実験

以上のように構成したシステムを用いて，学習に使用しなかった文章を入力テキストとして，音声と唇動画像の合成を行なった．音声合成部で得られた音素継続長を利用して，唇動画像 HMM の状態継続長モデルから各音素の状態継続長を計算する．この継続長に基づいて，唇動画像 HMM からパラメータ生成を行なった．

図 2 に「熱気のような」に対応する生成されたスペクトル系列と唇画像系列の一部を示す[1]．また，動的特徴の有無による生成されたパラメータの違いを図 3 に示す．実験結果から，動的特徴を考慮したモデルからは，実際の系列のように滑らかに変化するパラメータ系列が生成され，変換された画像系列においても，動的特徴量の有無による効果が確認された．

4. むすび

本論文では，HMM に基づく音声合成法を，画像ベースアプローチの唇動画像生成に適用した．さらに，これまでに提案してきた HMM に基づいた音声合成システムと組み合わせて，音声と唇動画像を同時に合成するシステムを構築した．固有唇を利用して特徴パラメータ次元を圧縮することで，音声と同程度のモデルパラメータ次元によって唇動画像のモデル化が可能であることを確認した．また，合成音声の音素継続長に合わせて唇動画像を生成することで，同期した音声と唇動画像を合成できることを確認した．

今後の課題として，唇だけでなく顔全体の動画像合成や，言語解析と組み合わせたシステムへの拡張が挙げられる．

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，音声合成部を提供していただいた名古屋工業大学博士課程 吉村貴克氏に感謝致します．

参考文献

[1] 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.

[2] 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.

[3] 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.

[4] 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), J66-A, 2, pp.122–129, 1983

[5] 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.

[6] 酒向 慎司, 徳田 恵一, 益子 貴史, 小林 隆夫, 北村 正 “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,”, 信学技報, PRMU99-157, Nov. 1999.

[7] M. A. Turk and A. P. Pentland, “Face recognition using eigennfaces”, Proc. of IEEE Computer Society Conf. on Computer Vision and Patter Recognition, pp.586–591, Maui, Hawaii 1991

[1] *http://kt-lab.ics.nitech.ac.jp/˜sako/lipsynthesis* に，生成された音声と唇動画像が置かれている