音声言語情報処理 30－1
（2000．2．4）

HMMを用いた入力音声からの自然な顔動画像生成

垣原 清次　　中村 哲　　鹿野 清宏

奈良先端科学技術大学院大学 情報科学研究科

email:{kiyotu-k,nakamura,shikano}@is.aist-nara.ac.jp

概要：入力音声から音声と同期した自然で現実感のあるコミュニケーションが可能な顔動画像の生成法を提案している．コンピュータを介した人間のコミュニケーションをより自然な形で実現できれば，コンピュータと人間のコミュニケーションの幅を飛躍的に広げることが可能である．我々は，以前に幾つかHMMを用いた音声からの唇動画像生成法を提案しており，特に後続音素の口形状を考慮することにより飛躍的に自然さを増すことに成功している．さらに本稿では，前後続音素の口形状を考慮した生成法を提案し，唇画像から顔3次元モデルへの拡張を行った．ここでは，顔表面3次元計測点に主成分分析を行い，主成分に対応した顔形状を予め作成することにより，入力音声からの自然で忠実な発話顔動画像の生成を実現した．

SPEECH-TO-FACIAL MOVEMENT SYNTHESIS USING HMMS

Kiyotsugu Kakihara, Satoshi Nakamura, and Kiyohiro Shikano

Graduate School of Information Science,
Nara Institute of Science & Technology

Abstract: This paper describes a talking face generation system with natural and communicative reality. If face movements are synthesized well enough for natural communication, a lot of benefits will be brought for the human-machine communication. We have already proposed a speech driven HMM-based lip movement synthesis, and also shown that the quality is drastically improved by considering succeeding visemes. This paper describes extension of our system to full face movement generation, and proposes a method considering both of preceding and succeeding viseme contexts. The experiments show the proposed method generates natural and accurate talking faces from audio speech inputs.

1. はじめに

近年，より自然な人間と機械のコミュニケーションを目指し，コンピュータ上に人間のように振る舞うエージェントを作成する研究が盛んに行われている．人間のような外観と応答性を持ったコンピュータエージェントの実現は，ヒューマンマシンインタフェースの大幅な改善に繋がると考えられる．エージェントの研究では，音声と同時に簡単な顔画像を提示してエージェントの内部処理状態を直感的知覚にさせたり，音声のバイモダリティを利用することにより音声認識を改善する研究が行われている．しかし，実際には，音声に対して自然で忠実な顔画像が合成されていないのが現状である．さらに，McGurk効果と呼ばれる人間の聴覚と視覚の感覚統合機能の解明により，音声と同時に呈示する発話顔画像には，音声と同期のとれた自然な動きの発話顔が必要であることが示されている．

これらのことから，著者らは，自然で発話内容が推測可能な発話顔動画像を合成することを目的として研究を行っている．発話顔動画像の合成には，テキスト情報から合成する方法と音声情報から合成する方法があるが，音声は同期した自然な顔画像得るための情報（韻律，継続時間長，感情）を多く含んでいるために音声から合成する方法を用いた．音声を用いれば，将来的には発話の顔のみでなく，感情に対応した表情も表現可能となる．

顔動画像合成には，レンジファインダで計測した顔3次元データを用いて任意の発話顔を生成しようとする研究[1]や，心理学分野で表情の表現のために利用されるFACS(Facial Action Coding System)の基本ユニットであるAU(Action Unit)を用いた研究

[2] などが報告されている．口形状は直接音声と対応するために，顔動画像合成の中でも特に唇動画像合成の研究が盛んに行われている [3]．この分野では，唇のパラメータからいかにして滑らかな唇を表現するかという画像側からのアプローチが重要な研究課題となる．

本稿では，顔表面 3 次元位置計測点に主成分分析 (以下，PCA:Principal Component Analysis) することにより顔合成で使用するパラメータ (以下，顔パラメータ) を決定した．また，顔パラメータをより忠実に表現するために顔動画像生成には 3 次元モデルを使用しており，基本顔形状から顔パラメータに対応する顔形状へ形状モーフィングにより変形することで発話顔形状合成を行っている．本手法は，唇周辺の滑らかな動きを含む発話顔形状の合成が可能である．

他方，テキストや音声から唇，顔動画像を合成する研究として，これまでにベクトル量子化 (VQ)[4] やニューラルネットワーク (ANN)[5]，隠れマルコフモデル (HMM) を利用した方法 [6, 7, 8, 9] が報告されている．我々は，以前に幾つかの HMM に基づく唇動画像合成法を提案して，コンテキストを考慮した音素認識が行える HMM を利用した手法が他の手法より有効であることを確認している．本稿では，HMM を利用した前後続音素の口形を考慮した顔パラメータ生成法を提案する．

以下では，音声と同期した発話顔動画像合成システムを構築すると共に，実際の発話時に近い自然で滑らかな顔動画像が合成可能なことを示す．

2. システム構成

音声から顔画像への変換とは，テキストなどの段階を踏まずに音声から直接発話時の顔形状や表情などを生成するものである．これは，入力音声から音声パラメータを抽出し，このパラメータに何らかのアルゴリズムによってその音声に同期した顔画像を規定する顔パラメータへと変換することと換言できる．図 1に，入力音声から顔画像合成法のブロック図を示す．モジュール (1) は，OPTOTRAK で収

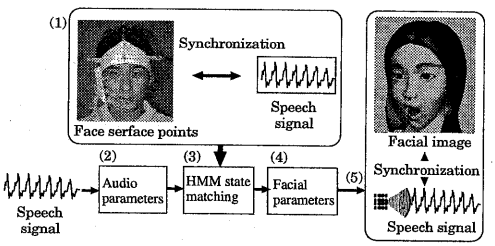

図 1: システム構成

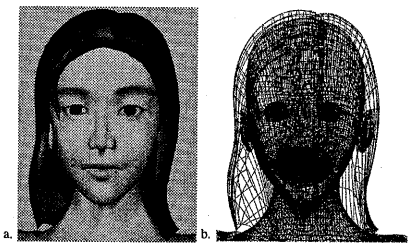

図 2: 顔 3 次元モデル (**a.** テクスチャマッピングモデル．**b.** ワイヤフレームモデル)

録した，音声と顔表面 3 次元位置計測点の同期したデータベースである．モジュール (2) でパラメータ化した入力音声をモジュール (3) で HMM に基づく変換法で音声パラメータから顔パラメータ (4) への変換を行なう．モジュール (5) では，顔パラメータから入力音声とそれに同期した 3 次元顔動画像を合成して出力する．

3. 顔パラメータから顔動画像生成

3.1. 顔 3 次元モデル

顔動画像合成には，顔 3 次元モデルを Alias Wavefront 社の Maya2.5 上で作成した．この顔 3 次元モデルは NURBS(Non Uniform Rational B-Spline) 曲面 6400 点の Primitive 球形状から一方を口道，他方に胴体を極にして一体整形されている．また，歯，口蓋，舌は別にポリゴンで作成した．顔 3 次元モデルを図 2に示す．

3.2. データ収録

顔に赤外線を発するマーカを貼り，CCD カメラにより 3 平面の交点からマーカの 3 次元位置座標を計測する OPTOTRAK(図 3:a) を用いて収録を行った．OPTOTRAK は，音声の分析周期に合せて 100Hz，実際の計測距離において計測誤差± 0.15[mm] という高レートかつ高精度な計測が行える．マーカの張付位置を図 3:b に示す．顔半面を計測対象とし，唇外側輪郭の周り 5 点，頬 7 点，顎 2 点で頭部のマーカ 4 点は原点同定に用いた．

3.3. 分析

顔表面位置計測点に PCA を行うことにより，顔パラメータの生成と次元数の削減を行った．顔表面位置点数が N の場合，各フレームの顔表面 3 次元位置は，x, y, z座標からなる $3N$の列ベクトル fで表現できる．Fを Kフレームの fを列に含む行列と

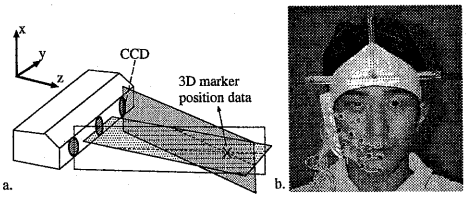

図 3: 顔表面の 3 次元位置計測 (**a.** OPTOTRAK **b.** 顔表面と頭部のマーカ位置)

すると,

$$F = [f_1, f_2, \cdots f_K]. \quad (1)$$

ここで，基本顔形状μ_fを口を閉じた状態の顔形状の列ベクトルと定義する．そしてFからμ_fを引いた値を基本顔形状からの差分行列ΔFとすると，

$$\Delta F = [\Delta f_1, \Delta f_2, \cdots \Delta f_K] \quad (2)$$

$$= [f_1 - \mu_f, f_2 - \mu_f, \cdots f_K - \mu_f]. \quad (3)$$

Fの主成分は，Fの分散共分散行列 C_fをユニタリ行列と対角行列成分の積に分解することにより計算できる．

$$C_f = \Delta F \Delta F^t, \quad yielding \quad C_f = USU^t. \quad (4)$$

式 (4) の Uは C_fの固有ベクトルを列に含むユニタリ行列であり，Sは対角成分に固有値を持つ対角行列となる．ここで Uは，

$$U = [u_1, u_2, \cdots u_K]. \quad (5)$$

3.4. 顔パラメータ

Uは主成分であり，フレーム kにおける基本顔形状からの差分ベクトルΔf_kを式 (6) のように表現できる．

$$\Delta f_k = U\alpha_k, \quad (6)$$

α_kは K次元の列ベクトルで，以下のように計算されるΔf_kの各主成分に対する射影成分である．

$$\alpha_k = U^t \Delta f_k, \quad (7)$$

本稿では，第q主成分の重みを顔パラメータ $PCA\,q$と呼ぶ．

Δf_kの要素は，各主成分 qの最大値 $PCA\,q_{max}$と最小値 $PCA\,q_{min}$の間で変化する．Δf_kの分散は，2 つの方向 (plus, minus) を持っているために，基本顔形状からの差分ベクトル顔形状 (以下，差分顔形状) は，$PCA\,q_{max}$と $PCA\,q_{min}$に対応する 2 通り必要となる．したがって，$PCA\,q_{max}$に対応した差分顔形状を PCA_{q+}，$PCA\,q_{min}$に対応した差分顔形状を PCA_{q-}とする．$PCA\,q_{max}$と $PCA\,q_{min}$は，予め学習データから求めておく．

3.5. 顔パラメータから顔 3 次元モデル合成

基本顔形状として標準的な顔 3 次元モデルを作成する．本稿では，収録データに対応した顔形状の合成にあたり，基本顔形状から各主成分に対応する差分顔形状を予め作成し，各次主成分の割合を固有値の大きさから求め，基本顔形状から各成分の差分顔形状へ形状モーフィングすることにより生成する．実際には，差分空間でなく実顔空間のモーフィングを行っている．簡単のため，以後，差分顔形状 $PCA\,q_{max}$と $PCA\,q_{min}$に基本顔形状μ_f を加えた顔形状を，それぞれ PCA'_{q+}, PCA'_{q-}とする．

3.6. 差分顔形状の生成

OPTOTRAK の収録データは点であり，顔 3 次元モデルは曲面である．従って，何らかの方法により顔表面 3 次元位置計側点から 3 次元モデル顔の顔表面を変形する必要がある．本稿では，計側点の影響範囲を決定しクラスタ化した．また，距離によって減衰する重みをかけ，PCA の結果に基づき各計測点に対応するクラスタを移動することにより，差分顔形状を生成した．顎の動きと歯，口蓋の動きは，収録データから顎の回転中心軸を求め，顎の回転量によりクラスタを移動する．唇の周辺の形状は複雑であり，クラスタ化の変形では滑らかに表現ことが出来ないために，変形した後に手作業でスムージングを施した．一度，PCA に対応した差分顔形状を生成すると，任意の発話顔形状合成が可能である．

図 4に ATR 音素バランス 216 単語の収録データのうち 10 フレーム毎のデータに PCA を行った結果から生成した第 1 主成分から第 5 主成分までの顔形状 $PCA'_{q-}, PCA'_{q-}(q = 1, 2, \cdots 5)$ と各顔形状の寄与率を示す．

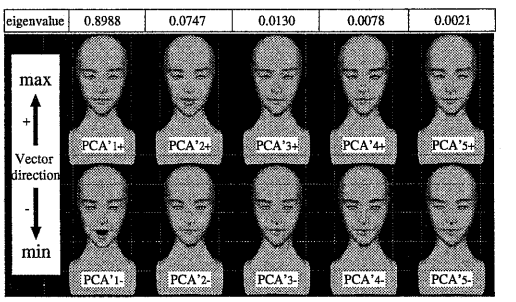

図 4: ATR 音素バランス 216 単語から生成した PCA に対応した顔形状と寄与率

4. HMMによる顔パラメータ生成

HMMによる音声に同期した顔パラメータ生成は，HMMの状態を単位として音声パラメータを顔パラメータに変換し，顔動画像を合成する方法である．この方法では，音素HMMの基づき，入力音声のフレームに対して尤度が最大となる音素HMMの状態を求め，その状態に対応する顔パラメータを連結することで音声に同期した顔動画像を出力する．このアルゴリズムは，音素HMMの状態に対応する顔パラメータを学習するアルゴリズムと，入力音声から状態系列を用いて顔パラメータに変換し顔動画像を合成する合成のアルゴリズムの2つにより構成される．ここでは，まずViterbiアライメントによる音素HMMの各状態への割り当てに基づいた顔パラメータ生成法であるHMM-V(HMM-Viterbi法)の学習と合成アルゴリズムについて述べた後に，今回提案する前後音素依存型HMM-V法を示す．

4.1. HMM-Vの学習のアルゴリズム

音素HMMの各状態に対応する顔パラメータは，学習データから同一のHMM状態に対応する顔パラメータの平均値を求めることで学習される．

[学習アルゴリズム]

1. 音声データと顔座標データを同期させて収録する．
2. 音声データと顔座標データの特徴抽出を行ない，音声パラメータと顔パラメータを得る．
3. 学習用音声データを音素HMMを用いて，発話全体で尤度最大となるフレームと状態の対応を決定する．
4. 全フレームのうち同じ音素HMMと状態番号をとるフレームを選択し，顔パラメータの平均値を求める．

4.2. HMM-Vの合成アルゴリズム

顔画像合成では，入力音声に対し，発話内容未知でViterbiアライメントを行ない尤度が最大となる音声フレームと音素HMMの状態に応じて対応する顔パラメータを取り出す．この過程を図5に示す．

[合成アルゴリズム]

1. 入力音声データをフレーム単位でパラメータ化する(音声パラメータ)．
2. 音素HMMを用いて，尤度最大となる発話内容未知のViterbiアライメントを求める．
3. 音素HMMの状態番号から，対応する顔パラメータを取り出す．

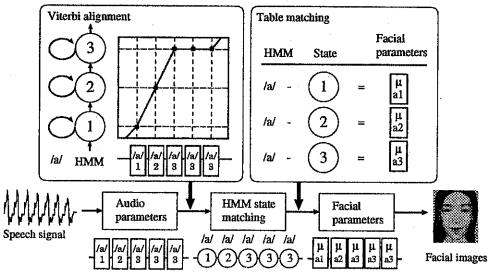

図5: HMM-Vによる音声からの顔パラメータ合成

4.3. 前後音素依存型HMM-V法(HMM-PSV)

HMM-Vで生成した顔パラメータでは，調音結合の強い音素と無音状態の誤差が大きい．これは，先行音素と後続音素の影響を考慮していないために起こる．この問題を解決するために，我々は，これまでに後続音素に依存した生成法であるHMM-SV(HMM-Succeeding Viseme法)を提案した[7]．そして，本稿ではより滑らかな顔パラメータ生成のために，先行音素も考慮した生成法であるHMM-PSV(HMM-Preceeding and Succeeding Viseme法)を提案する．

まず，顔パラメータ学習において，アライメントをとる段階で前後続音素に注目する．前後音素の影響は，PCAした際の第1主成分に強く影響することが分かっているため，$PCA1$の値によってクラスタリングした．この結果出来るクラスは口形状に対応するために，ここではVisemeと呼ぶ．図6に示すように，$PCA1$の値を小さい順にソートして3段階のクラスに分類している．先行音素HMMの第3状態と後続音素HMMの第1状態のVisemeに従ってクラスごとに各状態に対応した平均値をとる．このクラスタリングにより，顔パラメータの平均値は，HMMの各状態ごとに3(先行音素)*3(後続音素)=9個用意することになる．合成時にも入力フレーム毎に，音素，状態，前後続音素のVisemeを見て，平均値顔パラメータを出力する．

5. 発話顔形状合成実験

5.1. 実験条件

3次元顔モデルには前述のNURBS曲面一体整形モデルを用いる．ATR音韻バランス216単語の収録データの内10フレーム毎の顔表面位置データに対してPCAを行い，差分顔形状と顔パラメータを生成した．PCAの結果から得られる寄与率から，

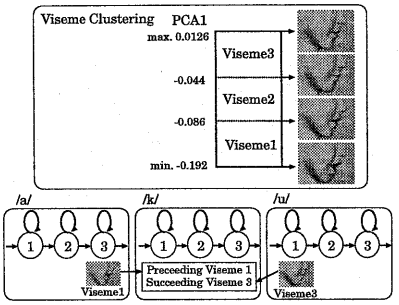

図 6: HMM-PSV で用いる Viseme のクラスタリング法

主成分の数を第 5 主成分までと決め，画像のフレーム数は 25 frame/sec とした．他方，音声データは 16kHz の音声データに 32msec 長のハミング窓をかけ，100Hz のフレーム周期で分析を行う．音声パラメータは，メルケプストラム係数 12 次元，メルケプストラムの差分係数 12 次元，差分 log パワー 1 次元の計 25 次元を使用している．音声認識に用いる HMM は，41 音素モデルと発話前，発話後の無声モデルの合計 43 個から成る．HMM は，Left-Right 型で 3 つの状態をもつ構造で，顔パラメータは，男性話者 1 名が発声した ATR 音韻バランス単語 216 単語により学習される．

5.2. 実験評価

結果は，マーカ 1 点あたりの距離誤差 Eで評価する．収録データのマーカ 3 次元位置を (x_o, y_o, z_o)，各手法 (HMM-V，HMM-SV，HMM-PSV) から生成したデータのマーカ 3 次元位置を (x_s, y_s, z_s) とすると，

$$E = \sqrt{(x_s - x_o)^2 + (y_s - y_o)^2 + (z_s - z_o)^2}. \quad (8)$$

表 1 に学習単語 216 単語とテスト単語 100 単語での平均距離誤差を示す．また，単語/triaezu/について，HMM-V で合成した顔パラメータを図 7 に，HMM-PSV で合成した顔パラメータを図 8に示し，実際に HMM-PSV で合成した顔動画像を図 9に示す．

表 1: 各手法の距離誤差

$E[mm]$	*closed*	*open*
HMM-V	2.31	2.40
HMM-SV	2.09	2.23
HMM-PSV	2.00	2.15

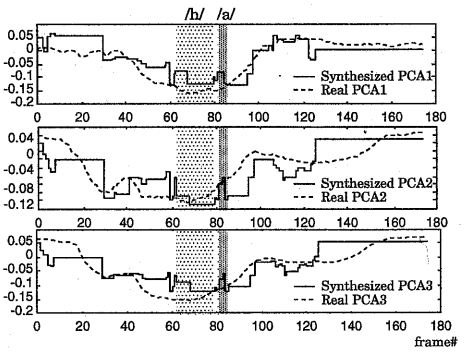

図 7: HMM-V で合成した顔パラメータ ($PCA1$, $PCA2$, $PCA3$)

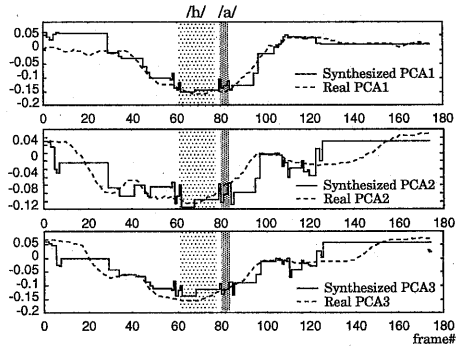

図 8: HMM-PSV で合成した顔パラメータ ($PCA1$, $PCA2$, $PCA3$)

このとき，HMM は */silB/-/sp/-/o/-/r/-/i/-/e/-/h/-/a/-/e/-/z/-/u/-/u:/-/u/-/silE/*と認識誤りしている．

距離誤差は HMM-V，HMM-SV，HMM-PSV の順で小さくなる．HMM-V と HMM-PSV の顔パラメータの比較では，音素間で明らかな改善が見える．このことから，HMM-PSV はより忠実で滑らかな顔パラメータを生成させることが可能となり，HMM-V 法と比較してより有効な手法であると言える．

6. まとめ

音声からの顔画像系列の合成において，顔表面 3 次元位置計測点に PCA を行ない，各成分の重みを顔パラメータとした．そして，3 次元顔モデルを使

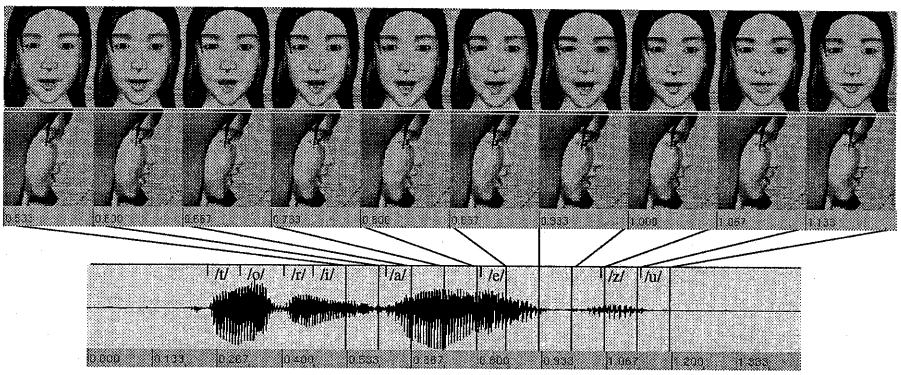

図 9: HMM-PSV で合成した顔動画像

用し，音声に同期た顔の動きを PCA の主成分に対応した基本顔形状からの差分ベクトル顔形状を作成して，基本顔形状からのモーフィングで合成することにより任意の発話顔形状を表現した．この手法は，計測した顔表面 3 次元位置計測点の移動量を顔 3 次元モデルに反映させる時に問題であった移動量への対応付けを，PCA に対応した顔形状を予め作成しておくことにより解決した．また，音声から顔パラメータの変換には，HMM-PSV 法を提案し，HMM-V 法より忠実さ，滑らかさの点で有効な方法であることを示した．

本手法を用いることにより，任意の音声から，別の 3 次元顔形状で別の人の発話特徴を持つ自然な顔動画像合成が可能なことから，ヒューマンマシンインタフェースの改善に留まらず，CG アニメーションの自動合成，顔と音声の組合わせが与える影響の心理学的利用を始め多くの分野での応用が期待できる．

今後，より忠実に顔計測点移動量を反映した PCA に対応する顔を作成すると同時に，主観評価実験を含め，この手法の有効性を確認して行く予定である．

7. 参考文献

[1] Kuratate, T., Hani, Y. and Eric V., "Kinematics-Based Synthesis of Realistic Talking Faces", AVSP'98, pp. 185-190, 1998.

[2] 長谷川 修, 森島 繁生, 金子 正秀, "「顔」の情報処理", 電子情報通信学会論文誌 (A), Vol. J80-A, No. 8, pp. 1231-1249, 1997.

[3] Lionel, R. and Christian, B., "A New 3D Lip Model for Analysis and Synthesis of Lip Motion in Speech Production", AVSP'98, pp. 207-212, 1998.

[4] Morishima, S. and Harashima, H., "A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface", IEEE Journal on selected. areas in Communications, Vol 9, No. 4, pp. 594-600, 1991.

[5] Lavagetto, F., "Converting Speech into Lip Movements", A Multimedia Telephone or Hard of Hearing People, IEEE Trans. on Rehabilitation Engineering, Vol. 3, No 1, pp. 90-102, 1995.

[6] Chou, W. and Chen, H., "Speech Recognition for Image Animation and Coding". ICASSP'95, pp. 2253-2256, 1995.

[7] Yamamoto, E., Nakamura, S. and Shikano, K., "Speech-to-Lip Movement Synthesis by HMM", AVSP'97, pp. 137-140, 1997.

[8] Yamamoto, E., Nakamura, S. and Shikano, K., "Speech-to-Lip Movement Synthesis based on EM Algorithm using Audio Visual HMMs", ICSLP'98, Vol.4, pp. 1275-1278, 1998.

[9] Masuko, T., Kobayashi, T., Tamura, M., Masubuchi, T. and Tokuda, K, " Text-to-Visual Speech Synthesis Based on Parameter Generation form HMM " , ICASSP'98, pp. 3745-3748, 1998.