音声言語情報処理 42－5
（２００２．７．１２）

実画像データベースを用いた発話顔画像合成の検討

白石 達也　　戸田 智基　　川波 弘道　　猿渡 洋　　鹿野 清宏

奈良先端科学技術大学院大学　情報科学研究科

あらまし　自然なコミュニケーションには音声情報とともに画像情報も重要な要素である．そこで本報告では，実画像を用いた発話顔画像合成手法を提案する．近年，音声合成技術ではコーパスベース方式が主流となっている．本手法では，コーパスに音声同期画像を用いることで画像へと応用することで，合成音声と完全に同期した発話顔画像合成を可能とした．本システムは３つのモジュールで構成され，顔位置探索モジュール・顔位置正規化モジュール・画像フレーム選択モジュールから成る．顔位置探索モジュールで眉毛・目・鼻・口領域を抽出し，それにより顔位置正規化モジュールでの正規化を可能とした．また，画像フレーム選択モジュールでは無音区間に対して viseme による分類を用いた画像フレーム選択を行った．得られた合成画像には，接続フレーム間の不連続性が残されており，今後その不連続性のパラメータ化を行う必要がある．

A Visual Speech Synthesis Method using Real Image Database

Tatsuya SHIRAISHI　Tomoki TODA　Hiromichi KAWANAMI
Hiroshi SARUWATARI　Kiyohiro SHIKANO

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

Abstract　Both speech and image information are important for natural communication. This paper describes a visual speech synthesis method using real image database. Recently, the corpus-based speech synthesis technique becomes in use. We apply that to the visual speech synthesis and make visual speech sequence. This system constitutes of three modules, face position searching module, face position normalizing module, frame selection module. First, eyebrow, eye, nose and mouth areas are extracted by the face position searching module. They are used normalize face position. In the frame selection module, ‘viseme’ classification is used for silence part. As the further study, we plan to parameterize discontinuity between neighbor frames and investigate appropriate method to reduce it.

1. はじめに

機械があたかも 1 個の人間のように振舞い，人間の顔や姿を表現し，音声言語で話し聞くような擬人化音声対話エージェントの実現を究極の目標として各種の開発が進められている[1][2]．コミュニケーションにおいて聴覚情報と並び視覚情報は非常に重要であり，自然な音声合成技術と共に自然な発話顔画像合成技術への期待が高まっている．

近年，携帯端末で電子メールの内容読み上げ，また盲人のための文書読み上げなどを目的としたテキスト音声合成（TTS: Text-To-Speech）技術が注目を浴びており，盛んに研究が進められている．現在，TTS はコーパスベース方式が主流であり，最適な音声波形素片をコーパスから選択して接続することにより，品質の高い音声を合成することができる[3]．一方，音声と同期した発話顔画像合成では三次元の顔モデルを利用した技術が一般的であり，効果的にモーフィングを使用することにより，小さい画像コーパスで表現することができる．しかしながら，そのコーパスの小さ故に合成発話顔に多様性が無く，またモーフィング技術は人工的な画像を生み出す[4][5].

そこで本研究では，実画像を用いたコーパスベースの発話顔画像合成を目指す．コーパスベース方式を採用することでコーパスを増加させることによる性能改善が期待できる．ただし，実画像を用いるため，コーパス中のそれぞれの画像における顔位置が異なるので，コーパス中から画像

を選択してそのまま連続再生しても自然な画像が得られないことが予想される．そのため，画像中の顔の位置情報を抽出し，顔位置に関する正規化等の処理が重要となる．

コーパスベースの発話顔画像合成法として，顔画像フレーム選択による合成法[6]が考えられるが，合成音声との完全な同期は困難であり，またシステム構築には多大な労力と時間を要する．そこで本報告では，より簡単な合成法として TTS システムからの出力を利用した発話顔画像合成法を提案する．本システムでは，あらかじめ音声と同期した発話顔画像データベースを使用する．合成時には TTS システムからの音声波形選択結果を利用し，それと同期した顔画像シーケンスを出力することで発話顔画像を合成する．しかしながら，TTS システムは音声合成システムであるため，無音区間に関する出力は一様である．発話者は，通常，発話に先立って調音器官を変形し，発話が終了した後で調音器官を元の形に戻すという傾向があり，本質的に調音器官の動きと発話区間とは一致しない．このため，無音区間でも顔画像は必ずしも静止していないことから，TTS システムで無音区間と出力される時間領域の顔画像をいかに自然に合成するかが重要となる．本報告では，この問題に対する解決策を述べる．

以下，2 章で本システムでの使用データベース，3 章で顔の位置情報を抽出するための顔位置探索法, 4 章で顔位置に関する正規化法, 5 章で画像フレーム選択について述べ，最後に 6 章においてまとめと今後の課題について述べる．

2. データベース

本システムはコーパスベース方式の発話顔画像合成であり，同期音声付き発話顔画像データベースを基に構成する．本システムで用いたデータベースについて表 1 に示す．収録は遮音室で行い，音声は接話型マイクで収録した．

3. 顔位置探索法

本探索法は，人物がカメラに対して正立した正面画像を対象として，色情報・顔の左右対称性・空間的な特徴を利用し，更に動画像を前提としたものである．以下にそれぞれの情報の利用法について述べる．

3.1 色情報

ここでは，人間の肌色情報を利用して顔肌領域をおおまかに抽出することを目的とする．色分布解析に用いる指標として，照明条件の変化による肌の明るさの違いや陰影などに対してロバストであることが望ましいため，HSV 表色系における H(色相)成分と S(彩度)成分を利用する[7]．HSV 表色系とは色の心理量を基準として色相(Hue)，彩度(Saturation), 明度(Value)の各属性に分解して表現するもので，色相は色の違いを区別する属性，彩度は色の鮮やかさを示す尺度，また明度は色の明るさを示す尺度となっている．ここで文献[8]より，色相と彩度の両分布とも正規分布に近似でき, 顔肌領域の確からしさ(尤度)を表す指標を色相と彩度を引数とした二次元正規分布の確率密度関数で表すことができる．色相と彩度の正規分布は次式で表される．

$$\begin{cases} N(\mathbf{\mu},\Sigma) = (2\pi)^{-\frac{1}{2}}|\Sigma|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\mathbf{x}-\mathbf{\mu})^T\Sigma^{-1}(\mathbf{x}-\mathbf{\mu})) \\ \mathbf{x}=\begin{bmatrix} x_h \\ x_s \end{bmatrix}, \mathbf{\mu}=\begin{bmatrix} \mu_h \\ \mu_s \end{bmatrix}, \Sigma=\begin{bmatrix} v_{hh} & v_{hs} \\ v_{sh} & v_{ss} \end{bmatrix} \end{cases} \quad (1)$$

x_h, x_s ：各画素における色相・彩度

μ_h, μ_s ：色相・彩度の平均

$v_{hh}, v_{hs}, v_{sh}, v_{ss}$：色相・彩度の共分散

色相と彩度の平均・共分散は，男性 6 名の顔肌領域を抽出し, その中のランダムな 5000 画素から算出した値を用いた．

発話者	男性 1 名
発話内容	ATR 音素バランス 503 文
収録方法	カメラに対して 正立した正面顔画像
画像収録機器	NAC HSV-500C3
画像フレームレート	125[frame/sec]
音声サンプリングレート	48[kHz]

表 1　音声付顔画像データベース

また原画像を図 1 として，これらから得られる顔肌領域の尤度を画像化したものを図 2 に示す．この画像は尤度を画素の明るさで表したものであり，明るいほど尤度は高いことを示す．

3.2 顔の左右対称性

ここでは，顔の左右対称性を利用して，前項で得られたおおまかな顔肌領域中から顔の中心線を検出することを目的とする．

人間の顔器官には，目・眉毛・耳等のように左右対称になっている部位が多く存在し，また鼻・口等の器官も中心線を引くことにより，ほぼ左右対象に同様の動きをする．これらのことから，顔肌領域の画像を左右に二分する y 軸方向(垂直方向)の直線を設けて，x 軸方向(水平方向)に走査させ，その直線により二分された顔の左右画像に対して左右対称性を考慮に入れた相関係数を算出し，その値が最も高くなる直線を顔の中心線とする．顔の中心線検出のイメージを図 3 に，顔肌領域から中心線を検出した結果を図 4 に示す．

3.3 空間的な特徴

ここでは，顔の空間的な特徴を用いて目・鼻・口等の顔器官領域の探索を目的とする．

文献[9]により，図 5 のような顔の各器官の領域矩形が得られている．これは，30 人の基本顔画像からそれぞれの顔器官領域の位置関係を調べたもので，両目間の距離情報のみから各領域の大きさと位置関係が決定する．この顔器官領域矩形を利用して，眉毛・目・口の各領域におけるサブコスト関数 $C_{eyebrow}(x,y)$，$C_{eye}(x,y)$，$C_{mouth}(x,y)$ を設け，次式で示すコスト関数 $C(x,y)$ が最小となるところで顔器官領域を決定する．

$$C(x,y) = C_{eyebrow}(x,y) + C_{eye}(x,y) + C_{mouth}(x,y) \quad (2)$$

探索範囲は，前項で検出した顔の中心線上である．以下にサブコスト関数について示す．

図 1　原画像

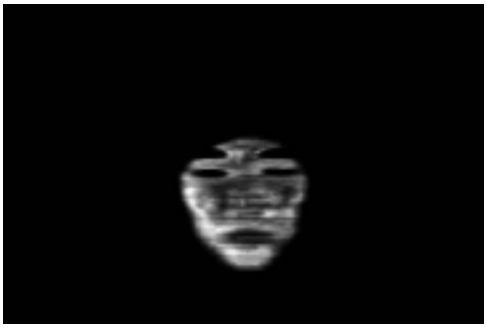

図 2　画像化した顔肌領域の尤度

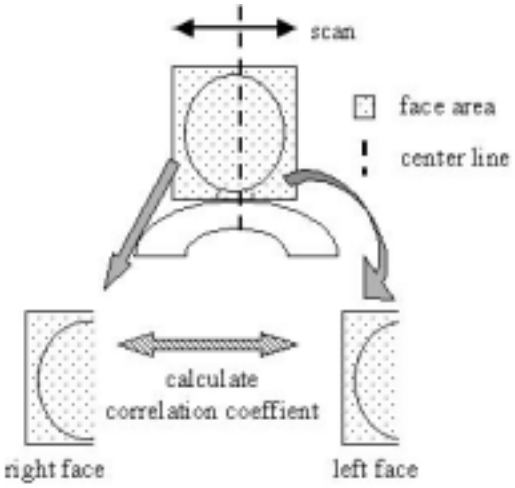

図 3　顔の中心線の検出

図 4　抽出した顔肌領域と中心線

ⅰ）$C_{eyebrow}(x, y)$：眉毛領域に関するコスト関数

このサブコスト関数は，眉毛領域であることの不適切さを表し，次式で表す．これは，眉毛領域は水平成分を多く含む顔器官であることによる．

$$C_{eyebrow}(x, y) = \sum_j \sum_i image\_across(x - i, y - j) \quad (3)$$

ここで $image\_across(i, j)$ とは，顔画像から水平成分を抽出した画像(図 6)の画素値を表す．

ⅱ）$C_{eye}(x, y)$：目領域に関するコスト関数

このサブコスト関数は，目領域であることの不適切さを表し，次式で表す．これは，眉毛領域と同様，目領域は水平成分を多く含む顔器官であることによる．

$$C_{eye}(x, y) = \sum_j \sum_i image\_across(x - i, y - j) \quad (4)$$

ⅲ）$C_{mouth}(x, y)$：口領域に関するコスト関数

このサブコスト関数は，口領域であることの不適切さを表し，次式で表す．これは，口領域は赤色成分を多く含む顔器官であることによる．

$$C_{mouth}(x, y) = \sum_j \sum_i image\_red(x - i, y - j) \quad (5)$$

ここで $image\_red(i, j)$ とは，顔画像から赤色成分を抽出した画像(図 7)の画素値を表す．

図 6,7 のいずれの画像も検出成分を明るさで表しており，暗い(画素値が小さい)ほど目的成分であることを示す．以上のコスト関数より，得られた顔器官領域探索結果を図 8 に示す．

また一般的に，動画像では隣接フレーム間の相関は非常に高く，隣接フレーム間で物体の位置や大きさ，角度等が大きく変動することはあまり想定されない．本探索法ではこの性質を利用し，2 フレーム目以降の顔探索は前フレームでの探索結果の近傍のみを探索することにより，探索時間の短縮と精度維持を図る．

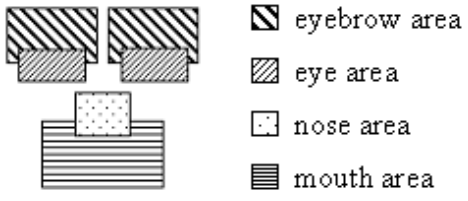

図 5　顔器官領域矩形

図 6　水平成分抽出画像　$image\_across(i, j)$

図 7　赤色成分抽出画像　$image\_red(i, j)$

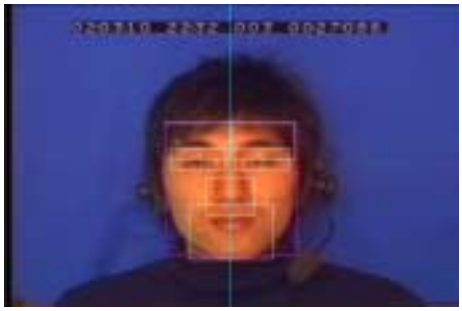

図 8　顔器官領域探索結果

4. 顔位置に関する正規化

本章では，合成発話顔画像の顔位置に関する正規化法について述べる．

4.1 正規化法

二次元平面顔画像の顔位置に関する正規化は，アフィン変換を用いて行う．変換前の座標を (x_1, y_1) とすると，変換後の座標 (x_2, y_2) は

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} a_{11} a_{12} \\ a_{21} a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \quad (6)$$

で表される．アフィン変換による正規化では，正規化を行う正規化点を画像中から抽出する必要があるが，画像中から一点を忠実に抽出することは困難であると考えられる．そこで本正規化法では，安定した正規化点を抽出するため，点ではなく領域を抽出することを考える．フレーム中の正規化領域を抽出し，2 フレーム間でその領域のずれを計測し補正することで顔位置に関する正規化を行う．

4.2 正規化領域

正規化領域は，各フレームにおいて比較的容易に抽出することが可能な領域で，かつ発話中に動きの少ない顔部位でなければならない．前述した顔探索結果より，眉毛・目・鼻・口領域は抽出が可能である．今回は，正規化領域の候補として目・鼻・口の領域を挙げ，発話中の動きが少ない領域を調査した．表 2 に目・鼻・口領域のそれぞれの基本画像(図 9)に対する相関係数の平均値と分散値を示す．この分散値が大きければ，発話中変動が激しい領域であることを示す．表 2 より，鼻領域は発話中変動が小さい部位であることが示された．また鼻器官は，水平成分の要素が多い顔器官の中で垂直成分を多く含む要素であるため，エッジ情報を利用した鼻位置探索を精度良く行えるという利点もある．以上より，画像コーパスをカメラに対して正立した人物の正面画像に限定すると，第 m, n フレームの顔画像の顔位置に関するずれ $shift\_face(m,n)$ は鼻の位置に関するずれ $shift\_nose(m,n)$ に近似することができる．

$$shift\_face(m,n) \cong shift\_nose(m,n) \quad (7)$$

以上より，鼻位置に関するずれを測定することにより，顔位置に関する正規化を行う．

5. 画像フレーム選択法

本画像フレーム選択法は，有音区間と無音区間で処理が異なる．本システムの概要を図 10 に示す．有音区間に関しては，TTS システム[10]で音声波形素片選択が行われ，それと同期した画像シーケンスを同時に出力する．これにより，音声と完全に同期した画像出力が可能となる．

	average	variance
eye	0.87064	0.00551
nose	0.94120	0.00047
mouth	0.69919	0.03323

表 2　各顔器官領域に対する
基本画像との相関係数の平均値と分散値

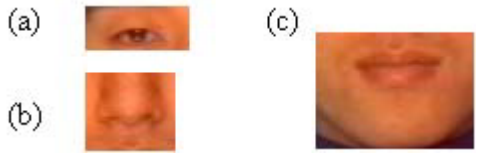

図 9　各領域の基本画像
(a)目領域　(b)鼻領域　(c)口領域

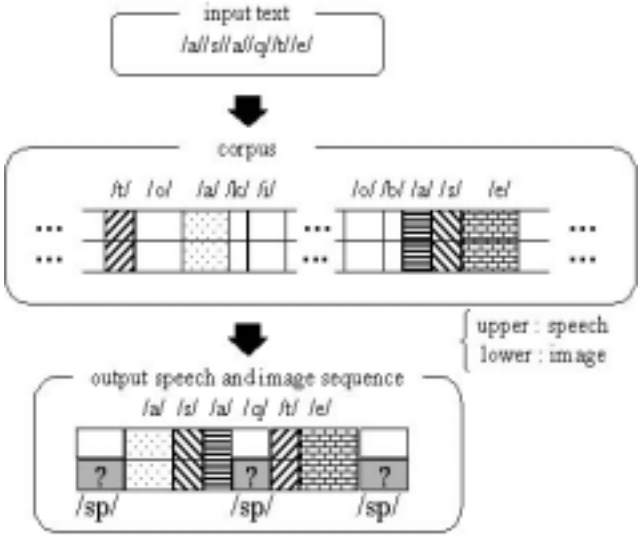

図 10　本システムの概略

しかしながら，無音区間に関する TTS システムからの出力は一様であるため，TTS システムから自然な顔画像シーケンスを得ることは困難である．そこで，TTS システムからでは画像出力が困難な無音区間に関する画像フレーム選択法を提案する．画像フレーム選択の探索空間は膨大であるため，口形素(viseme)を定義することで日本語子音に関する口形状の冗長性を削除し，更に，先行・後続の口形素環境を考慮に入れることで探索フレームの枝刈を行い，探索空間を縮退する．無音区間に関するフレーム選択のイメージを図 11 に示し，無音区間に関する画像フレーム選択の尺度として，口形素環境，口形状に基づく手法に関して以下で述べる．

5.1 口形素環境

異なる音素であっても口形状の変化は同一のものが日本語子音には存在し，画像フレーム選択する際に冗長性が高い．そこで，日本語音声の音素(phoneme)に対応する画像の口形素(viseme)を定義し，冗長性を削除する．口形素とは，ある音素を発しているときの唇の形状を表したもので，一般的に音素と口形素の対応はフレームレートや話速によって異なる[11]．しかしながら本報告では，30[frame/s]でごく一般的な話速での viseme クラスを想定し，筆者がヒューリスティックに決定した子音音素に対する viseme クラスを表 3 に示す．この viseme クラスを用い，先行・後続の口形素環境を考慮に入れることで，膨大な探索空間を縮退する．

5.2 口形状に基づく画像フレーム選択

前節で縮退した探索空間の中で，唇画像の類似度を用いて画像フレーム選択を行う．TTS システムにより得られた有音区間の画像シーケンス中の唇画像と，ターゲット唇画像との類似度を算出して不連続性をコストとし，最適な画像シーケンスを選択する．類似度として，相関係数を用いる．基本唇画像とそれぞれの唇画像との相関係数を図 12 に示す．この図は，唇画像の類似度を相関係数が表現していることを示しており，2 フレーム間の不連続性を表す尺度として適していることがわかる．

6. まとめと今後の課題

本報告では，実画像データベースを用いた発話顔画像合成手法を提案した．本手法では，顔位置探索，顔位置に関する正規化，並びに TTS システムの出力を利用した画像フレーム選択が行われる．更に， TTS システムの出力が利用困難な無音区間に関する画像フレーム選択を行った．

phoneme	viseme
k, h, g	viseme_k
p, b, m	viseme_p
f, w	viseme_f
d, t, n	viseme_d
s, ts, ch, z, j, sh	viseme_s
y, ky, ny, hy, ry, gy, dy	viseme_y
py, by, my	viseme_py

表 3　子音 viseme クラス

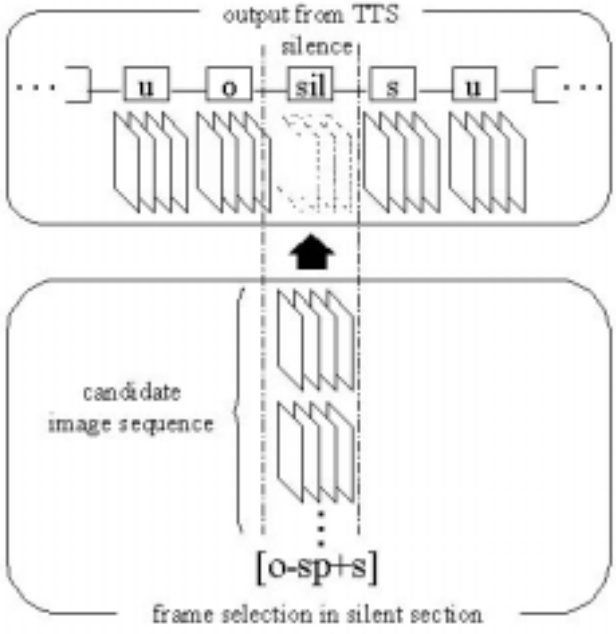

図 11　画像フレーム選択

しかしながら，出力された合成画像は十分に自然なものであるとは言い難い．これは，顔位置に関する正規化を行ったものの，接続フレーム間における画像の不連続性が未だ残されているためである．現在のシステムでは TTS システムの出力を利用し音声情報のみから画像フレームを選択しているので，このような不連続性が残されると考えられる．また，顔の角度を表すパラメータを用いていないことも原因として考えられる．

今後の課題として，両目と口が作る平面を表すパラメータを画像中から抽出することによって顔の角度を表現し，かつ音声情報を用いない新たな画像フレーム選択法を考慮することが挙げられる．また，今回はヒューリスティックに決定した viseme クラスであるが，定量的に最適な viseme クラスの決定を行う必要がある．そして合成発話顔画像に関して，自然発話画像等を用いた対比較実験など，定量的な評価を行う方法を検討する．

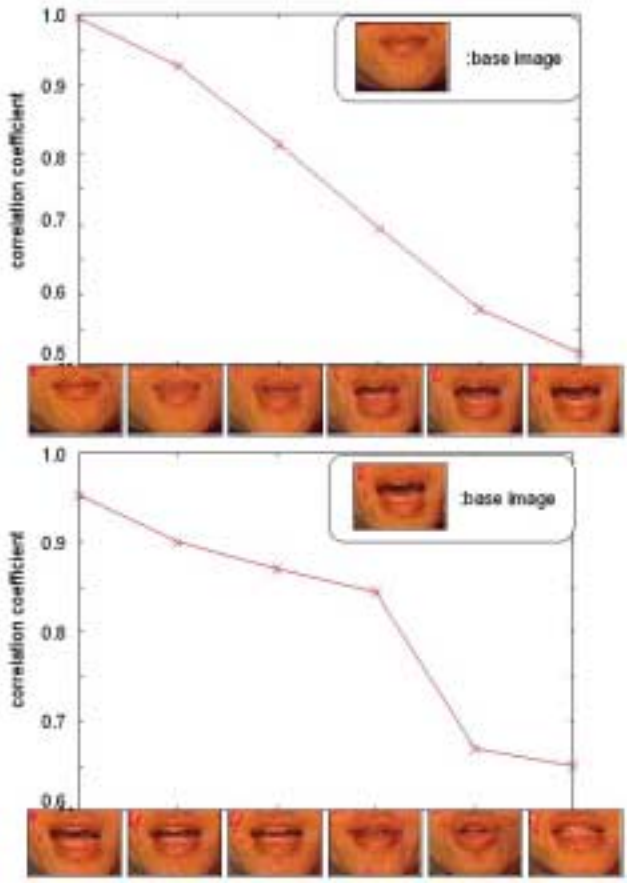

図 12　唇画像と相関係数の推移

参考文献

[1] 嵯峨山 茂樹，中村 哲，“擬人化音声対話エージェント開発とその意義，” 情報処理学会研究報告，SLP-33-1, pp.1-6, 2000.

[2] 森島 繁生，四倉 達夫，“擬人化音声対話エージェント開発とその周辺技術(3)対話における顔画像生成，” 情報処理学会研究報告, SLP-33-3, pp.13-18, 2000.

[3] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, “The AT&T Next-Gen TTS system,” Joint Meeting of ASA, EAA, and DAGA, Berlin, Germany, Mar. 1999. http://www.research.att.com/projects/tts/pubs.html

[4] M.Cohen and D.Massaro, “Modeling coarticulation in synthetic visual speech,” in Models and Techniques in Computer Animation, Springer-Verlag, 1993.

[5] Fabio Vignoli, “From speech to talking face: lip movements estimation based on linear approximators,” Proc. ICASSP, pp.2381-2384, Istanbul, June. 2000.

[6] F. J. Huang, E. Cosatto, H. P. Graf, “Triphone based unit selection for concatenative visual speech synthesis,” Proc. ICASSP, pp.2037-2040, Orlando, May. 2002.

[7] 高木，下田監修，“画像解析ハンドブック，” 東京大学出版会, ISBN4-13-061107-0, 1991.

[8] 船山 竜士，“複数の動的な網のモデルの協調による顔および顔部品領域の抽出，” 奈良先端科学技術大学院大学修士論文, NAIST-IS-MT9451096, 1996.

[9] 原島ほか，“感性擬人化エージェントのための顔情報処理システムの開発研究成果報告書，” イメージ情報科学研究所, 1998.

[10] 戸田 智基，河井 恒，津崎 実，鹿野 清宏，“日本語テキスト音声合成における音素単位とダイフォン単位に基づいた単位選択，” 信学技法, SP2001-120, pp.45-52, Jan. 2002.

[11] J.J.Willianms, J.C.Rutledge, D.C.Garstecki, and A.K.Katsaggelos, “Frame Rate and Viseme Analysis for Multimedia Applications,” Proc.IEEE First Workshop on Multimedia Signal Processing, pp.13-18, Princeton, NJ, June23-25, 1997.