グラフィクスとCAD 100−14
（2000. 9. 8）

主成分を用いたヘッドマウントカメラからの唇抽出とアニメーション応用

倉立 尚明†, Christopher Atkeson‡†, Eric Vatikiotis-Bateson†
kuratate@isd.atr.co.jp, cga@{cc.gatech.edu, isd.atr.co.jp}, bateson@isd.atr.co.jp

†(株) 国際電気通信基礎技術研究所　先端情報科学部
〒 619-0288 京都府相楽郡精華町光台 2-2
‡Georgia Institute of Technology, College of Computing
801 Atlantic Drive Atlanta, GA 30332-0280 USA

これまでに我々は、様々な口形状の三次元顔形状から得た主成分を用いることにより発話顔アニメーションを少ないパラメータで実現できることを示してきた。今回新たな試みとして、ヘッドマウントカメラから得た口周辺画像に対し、ＨＳＩ空間において色情報を元に唇候補領域を抽出し、それをあらかじめ求めた唇の外側輪郭形状主成分の線形合成で表現することにより、比較的高速に唇の外側輪郭形状パラメータを抽出することができた。そして、それをもとに他のＣＧキャラクタへの音声同期アニメーション生成を行ったので、それらに関して報告する。

Principal components based lip countour extraction from head-mounted camera and cross-subject facial animation

Takaaki Kuratate†, Christopher Atkeson‡† and Vatikiotis-Bateson†
kuratate@isd.atr.co.jp, cga@{cc.gatech.edu, isd.atr.co.jp}, bateson@isd.atr.co.jp

†ATR International, Information Sciences Division
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
‡Georgia Institute of Technology, College of Computing
801 Atlantic Drive Atlanta, GA 30332-0280 USA

Previously we demonstrated that talking heads animation could be represented by a small set of parameters by the linear combination of principal components which were extracted from various 3D faces of a single subject with different mouth postures. In this paper, we will present our new trial to extract lip motion from video and to map that motion to another computer graphics character. First, we analyze video data that was captured by a small head-mounted camera and roughly extract lip region by a simple segmentation and labeling scheme in an HSI color space. Then we represent outer lip contours by the linear combination of the principal components of outer lip shapes which were obtained at the pre-analysis stage. The linear combination coefficients are mapped to another CG character and we can easily get a speech synchronized facial animation.

1 はじめに

近年のインターネットの急速な普及により、これまでのラジオ、テレビのような単一のメディアでの一方向・受け身型の情報伝達から、複数メディアによる双方向・インタラクティブな情報伝達が盛んとなってきている。これに伴い、様々な情報のメディア変換、情報の効率的な蓄積、そしてインタラクティブ化も重要な課題となりつつある。

このようなことから、ジョージア工科大 Atkeson 教授は講義内容を複数ビデオカメラにより記録し、それらの画像情報から教室内の三次元構造の再構築、講演者の教室内の位置や四肢の動き、講演者の表情などの抽出と三次元モデルを生成することにより、従来の録音や単一のビデオカメラからのビデオ録画などの単調な講義記録ではなく、VRML などを利用した三次元的な仮想講義やインターネット上でインタラクティブに閲覧可能な形式での情報化を試みている。

今回は Atkeson 教授との共同研究として、これまでに我々が研究を行ってきた主成分に基づくアニメーション生成と [1]、他の話者への顔面運動のマッピング [2][3] を応用し、講演者の発話に伴う唇形状の動きを少ない情報量で蓄積もしくはリアルタイムでのデータ伝送を目的として数値化し、また他の CG キャラクタへその動きを移し代えるための実験を行ったので、これらについて報告する。

2 ビデオ画像からの唇輪郭抽出

2.1 講義内容のビデオ撮影

講義内容の記録には、講演者自体が装着するヘッドギアに固定された 4 台のカメラ (ヘッドマウントカメラ) と、教室内に設置された 7 台のカメラの合計 11 台のカメラを用いている。

4 台のヘッドマウントカメラは、口周辺・左目・講演者前方・講演者下方 (足元) をそれぞれ撮影し、講演者が装着するベストに組み込まれた Digital Hi8 ビデオにより個別に記録される。

教室内の 7 台のカメラのうち、4 台は講演者の上部の天井に固定された小型カメラで、講演者の講義中の位置を真上から撮影し、4 分割ビデオ入力装置により単一の Digital Hi8 ビデオに記録される。それ以外の 3 台は教室の天井前方に 1 台、後方に 2 台固定され、後方からの広角固定撮影を行っている 1 台を除き、残りの 2 台はカメラに組み込まれている色領域追跡アルゴリズムにより講演者が外側に着ているオレンジ色のコートを追跡している。

これらのビデオは音声トラックをもとに同期を取り、同一時刻での多視点情報を統合することにより、講演者の教室内における位置、胴体や四肢の動き、頭部の向き、口の動き、まばたき、視線方向等の情報を抽出し、最終的には CG キャラクタなどの他の話者モデルへの置き換えや VRML などによる三次元化などが可能な情報を抽出することと目標としている。今回は、この中の口周辺領域のビデオのみを用いて唇領域の抽出を行い、唇運動を CG キャラクタへの置き換えを試みたので以下にこれらについて解説する。

2.2 唇領域抽出と主成分表現

我々が用いたヘッドマウントカメラから得られる画像は、額部分を主として固定しているため、眉が動くような激しい感情表情や発話表現が現れない限りは頭蓋骨に対してほとんど定点観測となる。このため、唇領域についても現れる形状パターンについてもモデルを作ることが容易となる。

また、われわれの顔面運動のマッピングシステムでは顔面運動の主成分をもとに動きの投影を行っているため、唇形状についても同様な主成分による形状表現を得られた方が後段への処理において都合が良い。そこで唇形状についても主成分表現を求めることとした。

そこで、今回は図 1a に示すような単純な HSI 空間における色判別に基づいて唇領域を分離し、主成分分析および合成が容易にできるようなベクトルとして唇輪郭形状を求めることとした。

まず、明らかに唇領域ではありえないような領域に対する処理を省くため、入力画像に対しておおまかな処理対象領域を設定する。これは特にリアルタイム処理を目的とした場合、今回用いたようなヘッドマウントカメラ画像においてはわずかながら処理を高速に行うことが可能である。

そして、HSI 空間に変換した処理対象領域に対し、単純な閾値判定により唇領域候補画素を抽出する。しかしこの閾値判定ではノイズなどの影響が大きく、特殊なメークをした場合であっても安定した結果

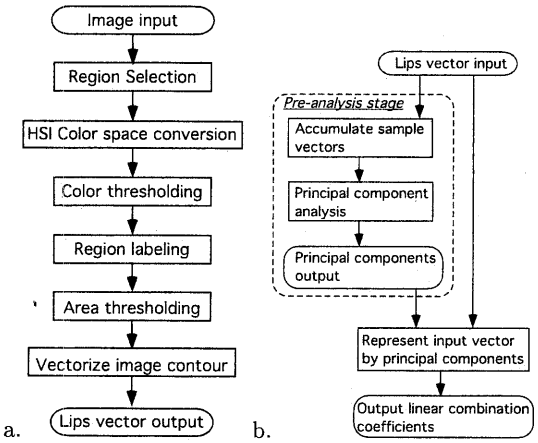

図 1: 本報告における唇抽出アルゴリズム (a) と入力唇ベクトルの処理アルゴリズム (b)

を常に得ることはできない。そこで、次に連結したピクセル領域をラベル付けし、これらの面積を用いた閾値判定を加えることにより、より安定した唇領域を得ることとした。HSI 空間および面積の閾値の設定はあらかじめサンプル入力画像に対して最適な値を求めていたものを用いている。

次に画素レベルで得られた唇領域を、輪郭形状として主成分分析を行うためにベクトル化を行う。ここでは、得られた唇領域を含む矩形画像領域を上下方向から探索し、画素単位での上側および下側での唇外輪郭座標を求め、これらの間を直線で結ぶ輪郭線に対し、あらかじめ定めた頂点数 N に適合するように構成点を内挿もしくは外挿し、得られた頂点の X,Y 座標をまとめて成分数 2N の唇輪郭ベクトルとしている。

入力画像から唇輪郭ベクトルへの唇形状の抽出ができた後は、図 1b に示すように前処理段階では複数の入力形状から主成分分析を行い、実際の処理ではこの得られた主成分の線形結合により入力唇形状を表現することができる。

2.3 唇輪郭の主成分分析

ここで、K フレームの画像から得られた唇輪郭ベクトルをフレームごとに列ベクトル $\boldsymbol{f_1}, \boldsymbol{f_2}, \ldots \boldsymbol{f_K}$ とし、各唇輪郭ベクトル $\boldsymbol{f_k}$ を平均 $\boldsymbol{\mu_f} = \frac{1}{K}\sum \mathbf{f_k}$ からの差 $\boldsymbol{f_{m_k}} = \boldsymbol{f_k} - \boldsymbol{\mu_f}$ として表現し、これらをまとめて前処理段階での分析用行列 $\boldsymbol{F_M}$ とする。

$$\boldsymbol{F_M} = [\boldsymbol{f_{m_1}}, \boldsymbol{f_{m_2}}, \ldots \boldsymbol{f_{m_K}}]. \tag{1}$$

この分析用行列 $\boldsymbol{F_M}$ を用いて、共分散行列 $\boldsymbol{C_f}$ は以下のように表すことができる。

$$\boldsymbol{C_f} = \boldsymbol{F_M}\boldsymbol{F_M}^t \tag{2}$$

この共分散行列 $\boldsymbol{C_f}$ に対して特異値分解 (singular value decomposition) を行うことにより、各固有ベクトルを $\boldsymbol{F_M}$ の線形独立な主成分として求めることができる。

$$\boldsymbol{C_f} = \boldsymbol{USU}^t \tag{3}$$

ここで $\mathbf{U}$ は正規化されたユニタリ行列で各列が固有ベクトルを表し、$\boldsymbol{S}$ は対角行列で対角成分が固有値を表すものである。

今回の解析対象である唇輪郭のように、非常に似通ったデータの主成分分析結果では、高次主成分は寄与率が低く、無視できるものが多い。そこで、任意の入力 $\boldsymbol{f}$ に対して、$\boldsymbol{U}$ から上位 n 主成分の列ベクトルのみを取り出した $\boldsymbol{U_n}$ を用いて主成分の線形合成係数 $\boldsymbol{\alpha}$ を求めると以下のようになる。

$$\boldsymbol{\alpha} \sim \boldsymbol{U_n^t}(\boldsymbol{f} - \boldsymbol{\mu_f}). \tag{4}$$

図 2に、唇に青のメイクを施した講演者による実際の処理対象画像と、それらに対する閾値処理結果、ラベル化と面積判定による抽出結果、得られた輪郭ベクトル ($N = 200$)、そして輪郭ベクトルを上位 3 主成分により表現した結果を示す。

また図 3に英語話者の発話データ (K=527 フレーム) に対して主成分分析を行った結果の上位 3 主成分を示す。図の上段から順に第 1、第 2、第 3 主成分を示し、各段の左側が (平均値 - 標準偏差)、右側が (平均値 + 標準偏差) における形状を示している。それぞれの段において、参考のため平均値形状を灰色で示している。これら上位 3 主成分が固有値に占める割合はそれぞれ 58.6%, 36.0%, 2.7% と、これらだけで 97% 近くを占めていることとなる。第 4 主成分以降は全て 1% 未満である。この図 3より、第 1 主成分が主として下唇の動き、第 2 主成分が上唇の動き、第 3 主成分が唇の両端の動きであることがわかる。特に上位 2 主成分について見てみると、第 1 主成分が唇の幅につい

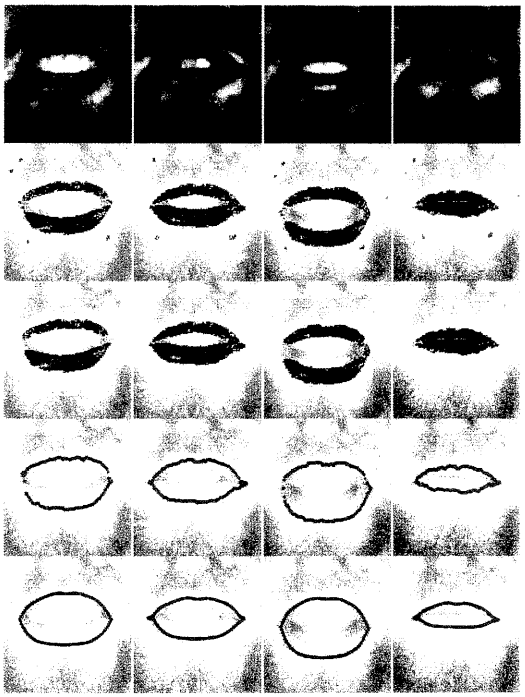

図 2: 唇抽出例: 上から入力画像、HSI 空間における色分離結果、ラベリングと面積判定結果、ベクトル化した抽出領域輪郭、上位 3 主成分による線形合成結果

ではあまり変化がないのに対して、第 2 主成分ではわずかながらではあるが幅に変化が見られる。発話データを観測してみると、主に上唇が大きく動く場合には唇を突き出すような動きが多く、この様な動きが第 2 主成分に現れ、第 1 主成分には下顎の上下運動に伴う下唇の開閉が大きく現れているのではないかと考えられる。第 3 主成分では両端の尖り方の変化が見られるが、これは領域抽出の際に照明等の影響により形状変化の大きい領域であり、実際の形状変化と共に領域抽出誤差も含まれているのではないかと思われる。

第 4 主成分以降は標準偏差内では形状変化を考察することは困難であるが、動きを強調して確認した限りでは領域抽出の際の画素単位での検出誤差ではないかと考えられる。このようなことから、今回は上位 3 主成分を用いて線形合成パラメータ抽出を行った。

ヘッドマウントカメラから得られる口周辺画像は、一定の照明条件ではここで用いた色情報から非常に簡単にかつ高速に唇領域が抽出できる。SGI 社の

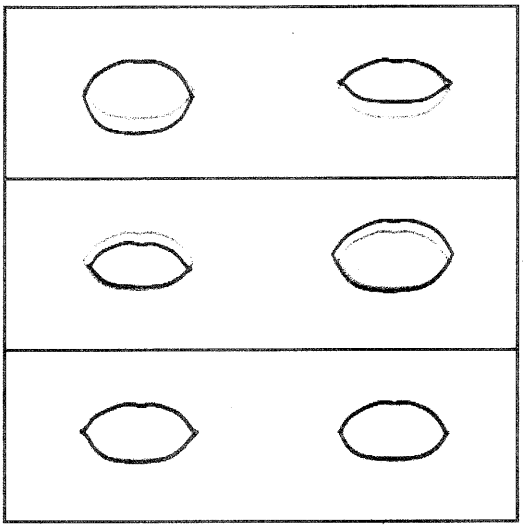

図 3: 主成分分析結果：上位 3 主成分における平均形状 (灰色) からの + 標準偏差 (左)、- 標準偏差 (右) に対する形状変化を示す

O2(CPU:R1000 180MHz) において、15Hz 程度でのメモリへの取り込みが可能なキャプチャプログラムに対して本アルゴリズムを適応したところ、約 12Hz 程度で上位 3 主成分による線形合成パラメータの抽出が可能であり、リアルタイム化への応用が期待できる。

しかし、実際に記録した講義においてはビデオ映像やコンピュータ映像をプロジェクタを用いて提示することが頻繁にあり、このため講演者がプロジェクタの投影光内を横切る場面が多々あり、これによる入力画像の色分布や輝度分布が大きく変化することが多く、フレーム間の相関なども加えた処理が必要である。

3 主成分による顔面運動マッピング

3.1 CG キャラクタの主成分抽出

次に顔面運動マッピングを利用して、得られた線形合成パラメータを他のＣＧキャラクタへの投影を行う [2][3]。まず、ＣＧキャラクタ側では母音や口を開ける、閉じるなどの特徴的な口形状を用意して、唇輪郭の場合と同様に分析用行列 $\boldsymbol{F}_{M_{cg}} = [\boldsymbol{f}_{1_{cg}}, \boldsymbol{f}_{2_{cg}}, \cdots \boldsymbol{f}_{K'_{cg}}]$ に対して主成分分析を行い、上位 n_{cg} 主成分 $\boldsymbol{U}_{n_{cg}}$ を求める。これにより、$\boldsymbol{U}_{n_{cg}}$ と線形結合係数 $\boldsymbol{\alpha}_{cg}$ により CG キャラクタの顔

形状 f_{cg} を合成することができる。

$$f_{M_{cg}} \sim U_{n_{cg}} \alpha_{cg}, \tag{5}$$

また元の形状 $F_{M_{cg}}$ は、完全ではないが $U_{n_{cg}}$ を用いて以下の式で近似することができる。

$$F_{M_{cg}} \sim U_{n_{cg}} A_{cg}, \tag{6}$$

$$A_{cg} = [\alpha_{1_{cg}}, \alpha_{2_{cg}}, ...\alpha_{K'_{cg}}]. \tag{7}$$

これは唇形状についても同様に A_{lip} が定義できる。

今回は犬のCGキャラクタを分析・合成の対象とし、顎・頬・上唇近辺に形状変化を与え、口を大きく開ける、自然に閉じる、噛みしめて閉じる、母音 /e/ を発声しているような形状のK'=4種類の形状から主成分を求めた。その結果、第1主成分として顎を閉じる／開ける (寄与率91.91%)、第2主成分として頬および上唇部分を上げる／閉じる (8.06%)、第3主成分として上唇の中央部分を上げる／下げる (0.03%) という結果を得ている。図4にこの3主成分の結果を示す。図の上段から順に第1,2,3主成分を示し、各段の左側が (平均値 - 標準偏差)、右側が (平均値 + 標準偏差) における形状を示している。第3主成分では標準偏差での変化は小さく判別しづらいが、この図では目と鼻の中点から舌に向かって真下に向かう部分で、上唇がかすかに上下方向に変化している。

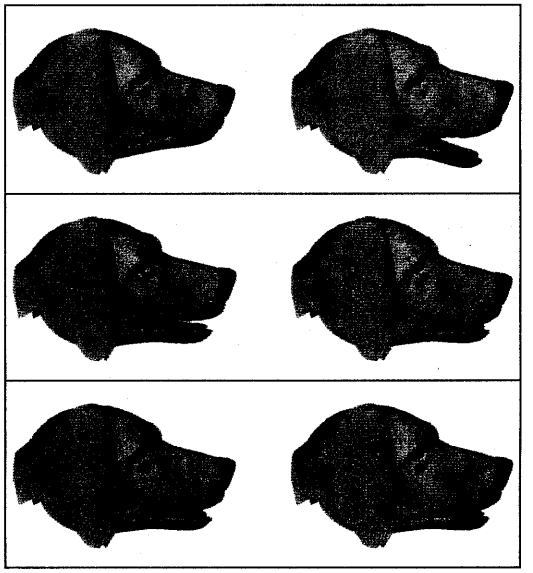

図 4: 四種類の形状から得た3主成分：左側が平均 - 標準偏差、右側が平均 + 標準偏差

3.2 顔面運動マッピング

次にビデオから求めた唇輪郭の主成分をCGキャラクタの主成分へマッピングを行う。このためには、唇輪郭の主成分を求めるために用いたデータとCGキャラクタの基本形状との対応と、主成分の対応関係が必要となる。そこで A_{lip} から A_{cg} への線形推定子 E_{map} を考えると、以下の関係が成立する。

$$A_{cg} \sim E_{map} X A_{lip} S = E_{map} A, \tag{8}$$

$$E_{map} \sim A_{cg} A^t (A A^t)^{-1}. \tag{9}$$

ただし、$A = X A_{lip} S$ である。ここで S は $K \times K'$ の基本形状の対応行列で、行方向を入力唇形状、列方向をCGキャラクタ形状に対応させ、お互いに基本形状が同じと見なせる行と列の交差成分を1、それ以外を0とする行列である。また X は $n_{cg} \times n$ の主成分の交換行列で、行方向をCGキャラクタの主成分、列方向を唇形状の主成分に対応させ、同じような主成分が存在する場合にその交差成分を1もしくは-1、それ以外を0とするものである。今回の例では X は $n_{cg} = 2, n = 3$ として第1、第2主成分のみを対応させ、特に口の開きが双方で反対であることから以下のように定義した。

$$X = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} \tag{10}$$

二種のデータにおける主成分分析と類似性を求めておけば、E_{map} はあらかじめ定義することができる。この E_{map} と、唇形状の主成分の線形合成係数 α からCGキャラクタを線形合成する係数 α_{cg} が以下の式で与えられる。

$$\alpha_{cg} \sim E_{map} I_X \alpha \tag{11}$$

ここで I_X は X の非0成分を1としたものである。

α_{cg} が与えられることにより、式 (5) に基づき唇形状に対応したCGキャラクタを合成することができる。図5に入力画像とCGキャラクタへのマッピング結果を示す。この図より、唇の画像上での状態に応じてCGキャラクタの口が開閉している様子がわかる。ただし、こののCGキャラクタのデータでは唇の突き

出しのような形状が含まれていないため、図の4段目にある、唇を突き出したような形状は、口の開き具合いが似通った3段目の場合と同じようなパラメータとしてマッピングされるため、結果は同じような口の開き具合いのみが再現されている。

また、基本的に今回の処理は全てフレーム毎で独立して処理しているため、唇輪郭抽出の際にフレーム間での検出精度の違いなどが多い場合は、生成されるアニメーションでもフレーム間の動きの飛びなどとして現れている。

現在はこれらのアニメーション映像はオフラインで生成しているが、専用のグラフィック生成サーバと、唇輪郭処理サーバとを異なる計算機上でそれぞれ処理すれば、実時間でのCGキャラクタへの動きの投影も実現可能である。

4 まとめと今後の課題

ヘッドマウントカメラからのビデオ画像から比較的高速に唇輪郭形状を主成分の線形合成係数として求め、それを顔面運動マッピングによりCGキャラクタの動きに移し代えることができた。

さらに今後はよりリアルな動き・形状変化を再現するため、唇の内側輪郭を含めた二次元唇形状や、口腔内情報をふくめた画像情報からより高精度な三次元顔形状を復元することが考えられる。

また、今回は複数視点からの講義記録ビデオのうち、唇領域を観測したビデオ画像のみを用いている。しかし、発話に伴う動きとしては頭部運動も重要な役割を果たしており、発話に応じた自然な頭部運動は音声の知覚に対しても影響が大きいことから、他のビデオ画像を利用して頭部運動も加えたアニメーション生成を試みることも検討している。

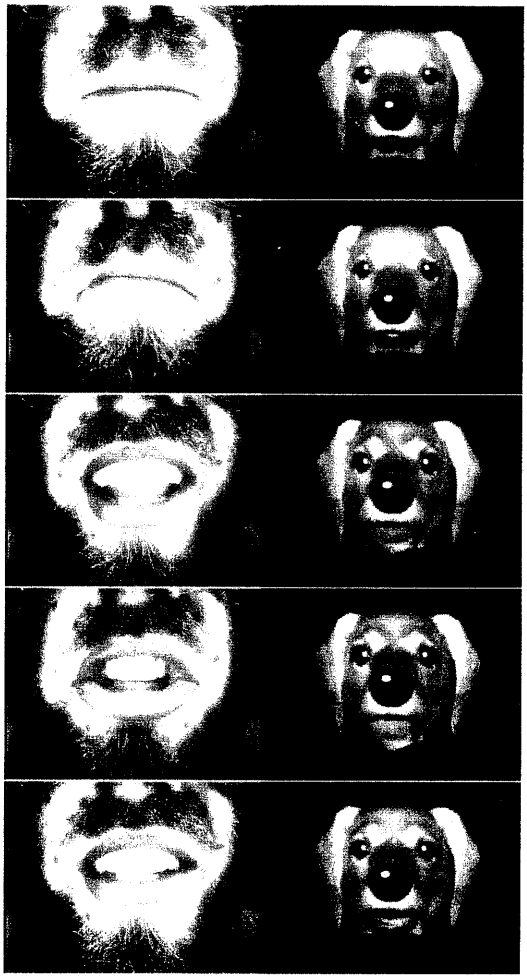

図 5: 入力ビデオ画像とマッピング結果

参考文献

[1] 倉立 尚明, Hani Yehia, and Eric Vatikiotis-Bateson. 主成分の線形結合による音声同期顔アニメーション. *Visual Computing* グラフィクスと *CAD* 合同シンポジウム, pages 115–120, 1998.

[2] 倉立 尚明 and Eric Vatikiotis-Bateson. 主成分分析に基づくアニメーション応用. 画像電子学会 *Visual Computing Workshop in Toba*, 1998.

[3] Takaaki Kuratate, Eric Vatikiotis-Bateson, and Hani Yehia. Talking face synthesis by facial motion mapping. *Speech Communications*, in press.