2004－HI－110　(4)
2004／9／10

モーションキャプチャシステムを用いたマルチモーダル音声コーパスの構築

四倉　達夫[1]　　森島　繁生[1,2]　　中村　哲[1]

[1] ATR音声言語コミュニケーション研究所
[2] 早稲田大学理工学部
Email: {tatsuo.yotsukura, shigeo.morishima, satoshi.nakamura}@atr.jp

あらまし:
本稿では音声、顔画像および発話時における顔器官の位置とその変化量を含むマルチモーダル音声コーパスの制作方法、およびデータの処理方法について述べる。発話用テキストはATR日本語バランス文とし、女性話者1名の発話をコーパスとした。変化量の計測には光学式モーションキャプチャシステムを使用し、発話者の顔上に多数のマーカを配置することで、顔画像情報のみでは獲得することができない顔位置の詳細かつ高精度の3次元データを収録した。さらに本稿では純粋な顔器官の動きを算出するため、アフィン変換を用い頭部の動きを除去し顔位置のみの情報を獲得する手法を提案する。またコンピュータ上に計測した変化量を発話アニメーションへ容易に再現させるため、顔器官の動きをメッシュで構成された顔オブジェクトへ割り当てる手法について述べる。

Construction of Audio-Visual Speech Corpus using Motion Capture System

Tatsuo YOTSUKURA[1], Shigeo MORISHIMA[1,2] and Satoshi NAKAMURA[1]

[1] ATR Spoken Language Translation Research Laboratory
[2] School of Science and Engineering, Waseda University
Email: {tatsuo.yotsukura, shigeo.morishima, satoshi.nakamura}@atr.jp

Abstract:
In this paper, we describe the construction and the processing method of the multi-modal speech corpus, which contain speech data, facial movie data and position and movements of facial organs. One female speaker uttered ATR Japanese phoneme balanced sentences. Measurement of the facial movements is done by an optical motion capture system. We captured high-resolution 3D data by arranging many makers on the speaker's face. Furthermore, we propose the method of acquiring the facial movements and removing head movements using affine transformation for computing displacements of pure facial organs. Finally in order to represent facial animation from this motion data easily, we describe the technique of assigning the facial polygon model.

1. はじめに

音声認識または音声合成に関する近年の研究の進歩は隠れマルコフモデル（HMM）に代表される大規模音声コーパスを用いた統計的アプローチに基づいており、さまざまな研究機関がコーパスを構築している。最近では視聴覚情報、すなわち音声のみならず動画像を同時に利用した音声認識、音声・顔画像合成に関する研究が盛んに行われるようになり、その為のマルチモーダルコーパスの整備も進められている[1]。特に顔画像合成に関する研究は、音声認識・合成で使用している統計的アプローチを利用した顔動画像コーパスを利用して Talking Head 生成を行う研究機関も多い[2]。この Talking Head はデータベース中の顔画像から再生成されるため、メッシュで構成された顔モデルを用いて、メッシュ形状を変形させ表情変化を行うモデルベースによるアプローチ[3]と比べると表情変化のバリエーションが制限されてしまう欠点があるが、コーパスが大規模かつ適切なデータであれば、顔モデルを使用しないため自然性が向上し、表現力が向上する。このように様々な研究分野においてもマルチモーダルコーパスに利用範囲は大きく、多くの需要が見込まれると考えられる。

さて、これら研究に用いられる際、音声処理において音声波形から F0、パワー、MFCC など音声特徴量を求めて分析を行うのと同様に画像処理においても画像特徴量を抽出する必要がある。画像特徴量としては顔器官の特徴量である、顔部位の変化量を得るためピクセル情報からオプティカルフローを求め顔器官の動き推定を行う手法[3]や唇領域のピクセル情報を用いる手法[4]などを用いて特徴量を得る方法が知られている。しかし発話時における顔器官の変化量を単一のビデオカメラ等によって撮像された 2 次元の画像情報から正確に得ることは非常に難しい。

さらに安価なデジタルビデオカメラを使用した際は安定したカラーバランスで撮影することは難しく、レンズや CCD により画像の歪みが生じる恐れがあり、特徴量を算出するときはこれらを考慮する必要がある。

また精度が良く、画像解像度の高いカメラを使用したとしても、3 次元の位置を知ることは難しく、発話時の頭部の位置がずれた場合の正規化を行うことも困難である。

頭部の動きは体全体の動き、頚椎による頭部の回転、移動が影響しているため 3 次元空間で非常に複雑な動きとなる。

そこで本稿ではマルチモーダル音声コーパスの構築を行うに当たり音声、顔画像のみならず正確な顔器官の変化量をコーパスに含めることを目的にし、著者らが行った撮影方法、データ処理方法について述べる。顔器官の変化量を測定するため、光学式モーションキャプチャシステムを採用し 3 次元計測を行った。高精度かつ高解像度に顔の形状を計測可能なレンジスキャナシステムも有効な手段の一つに挙げられるが、撮影時間に数秒要するため時間解像度が必要な本撮影には適さない。今回使用した光学式モーションキャプチャシステムは、最低毎秒 30 フレームの計測が可能であり（本計測では毎秒 120 フレームで撮影）、さらに発話者の顔の上に 130 点以上の直径 3~4mm 程度のマーカを配置して撮影を行うことにより、高精度かつ詳細な顔部位の変化量の獲得が可能となる。

本稿ではさらにマーカの動き情報からメッシュで構成された顔オブジェクトへ割り当てる手法についても述べる。この手法を用いることで各自制作した顔モデルを容易に発話や表情変形させることができる。さらにこれは 3 次元ソフトウェアの Plug-in として整備することによりコーパス整備以外に、モーションキャプチャデータからクリエータらがそのソフトウェアで制作した 3 次元キャラクタモデルの顔に表情を容易に割り当てることが可能である。

2. コーパス収録法

2.1. コーパス撮影プロセス

本稿でのコーパス作成過程を図 1 に示す。コーパスの収録にはバストアップの顔画像を正面から DVCAM カメラにて撮影した。同時に DAT デッキを使用して音声を録音した。さらにモーションキャプチャシステムを使用することで顔顔部位の動きを測定した。モーションキャプチャシステムは Vicon 社の Vicon612 を使用している。このシステムは光学反射式で高解像度・高速度の複数の赤外線カメラで撮影エリア内の高再帰性光学反射マーカの位置を撮影し、その映像から

ソフトウェアにてその 3 次元座標が算出可能となる。本撮影ではこのマーカを多数、顔皮膚に装着することで発話時の顔の動きを捉えることを可能とした。尚、モーションキャプチャシステムと DVCAM コーダのタイムコードはタイムコードジェネレータ（BetaCAM コーダを代用）から出力した VITC(Vertical‐Interval Time Code)信号を入力しているため同期したデータが収集可能である。

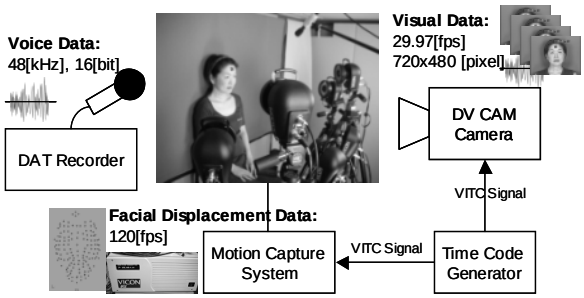


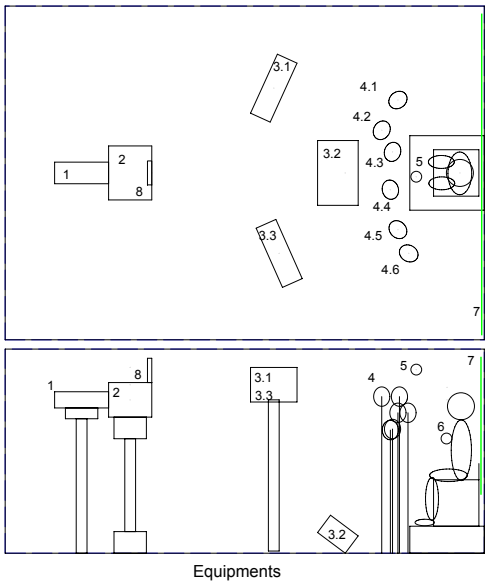
図 1　コーパス作成工程

2.2. 撮影環境および手法

撮影時の風景および機器のレイアウトを図 2、3 に示す。撮影場所は ATR 音声言語コミュニケーション研究所の防音室において行われた。撮影時、発話者の顔にセルフシャドーが起こらないよう、発話者の左右および正面ローアングルから照明を投射した。照明はフリッカーフリー、高出力、一定の色温度を保つ照明用蛍光灯を使用した。カメラフォーカスはマニュアルで合わせ（F=4.0、3m）、カメラ補正は色温度を 5000[K]に設定し、ノンインターレスにて撮影した。また、撮影後の画像処理を考慮し被験者の背後には緑色のクロマキースクリーンを配置した。

図 2　撮影の様子

収録したコーパスのテキストは ATR 日本語データベースの音素バランス文 B セット 503 文章および、音素連鎖バランス単語 216 個、単音節 101 個、数字・アルファベット 55 個で用いられているものを使用した。発話者はプロアナウンサー経験を有する女性 1 名とし、発話の前後に口を閉じるよう指示した。これら発話リストは事前に原稿として発話者に渡し練習するよう依頼し、さらに撮影時はテレプロンプタを用意し、なめらかな発話ができるようにした。テレプロンプタは発話者の視線をレンズ方向へ一定に保つためにも使用している。さらに発話者が自分の収録の様子が見ることができるよう、テレプロンプタ上部に液晶モニタを配置した。



Equipments

1: DV CAM Camera:Ikegami HL-DV7W
Lens: Canon IFXS J16a x8B4
Tripod: Sachtler IRS-C Sx12 VIDEO
2: Tele-Prompter: Canon CWP-10H
3: Lights: Lowel Fluo-Tec Caselite4 x 3
4: Motion Capture Cameras: Vicon MCam x 6
5: Microphone: Audio-technica AT4051a
(Small Diaphragm Cardioid Condenser Microphone)
6: Microphone: Audio-technica ATM14a
(Non-directional microphone)
7: Chroma Key Screen (Color Green)
8: LCD Monitor: SONY

- Preview Monitor Monitor: SONY PVM-20M4J
- Processing PC for MoCap : HP Pentium4 3.06GHz
- Motion Caputure System: Vicom612 Datastation
- Microphone preamplifier: TASCAM MX-4
- Time Code Generator (BetaCam): SONY BVW-76
- DAT Recorder: DAT TCD-D10 PRO2

図 3　機器一覧およびレイアウト

収録用マイクロフォンは発話者の上部および、胸部に配置し 2ch にて DAT を用いて収録した。モーションキャプチャ用のカメラは被験者を囲むように 6 台配置した。モーションキャプチャ用マーカの配置方法に関しては次節で述べる。

本撮影では音声、顔画像、顔部位の変化量、合計 3 つの情報を同時に撮影する。さらにモー

ションキャプチャの計測においては顔全体にマーカを配置、顔右半分にマーカを配置、額4カ所のみにマーカを配置、合計3パターン行い、いずれも全ての文章および、単語の発話を撮影した。

2.3. モーションキャプチャマーカの配置

顔の皮膚上にマーカを配置し、これらの動きをカメラによってトラッキングすることでの顔部位の動きを測定することができる。VICON社が推奨する一般的なマーカの配置は30点程度であるが、本稿では発話時での顔部位の詳細な変化量を必要とするため眉部17点、目輪郭部17点、鼻部7点、唇部17点、頬部36点、顔輪郭部15点、顎部10点、首部14点、額部4点、合計137点配置した。配置に要する時間は約1時間30分である。これらマーカの配置順序および配置方法は複数の発話者を測定するときにでも同様の配置ができるようルールを設けた。さらに発話時の口周辺の詳細な変化量を知る必要なため唇、頬、顎にマーカを多く配置しており、眉部と目輪郭部にも発話との瞬き関係を検証するため配置した。さらに額部にも被験者の顔が動いたときのデータ補正用に配置した。額自身の動きを抑制するため、厚手のテープを数枚重ねた物を額に貼りその上にマーカを装着した。本稿では3mmの半球、4mmの球状のマーカを使用した。なお、顔半分のマーカ配置のときは図4のように78個のマーカを配置した。

3. データの取り込み

3.1. 動画像・音声データの取り込み

収録された動画像データはキャプチャカード（Canopus DVRex-RT）およびソフトがインストールされたPCを用いてDV-CAMテープからキャプチャを行った。キャプチャ後のデータは720x480[pixel]、フルカラー（RGB=24[bit]）、29.7[frame/sec]のCanopus DV-CAM圧縮形式の動画像となった。さらに動画編集ソフトを使用して文章、単語単位に手作業で切り出しを行いファイル化した。音声データはDV-CAMの取り込み時に動画像と同時にサンプリングレート48[kHz]、16[Bit]のPCMデータがキャプチャできるが、DV-CAMコーダと比べDATレコーダの方が音声のA/D変換器の性能が高いため、DATデータを使用した。DATで収録されたデータはCanopus社のメディアコンバータを使用してPCへキャプチャ（48[kHz]、16[Bit]）を行う。ただDATレコーダにタイムコードの入力が無いため、DATデータのタイムコードはDV-CAMやモーションキャプチャデータと同期していない。その為、DV-CAMデータでキャプチャした音声データとDATデータとの相互相関を計算し、その最大で同期を取った。同期された音声データは動画像データの切り出した手法と同様に手作業でファイル化した。

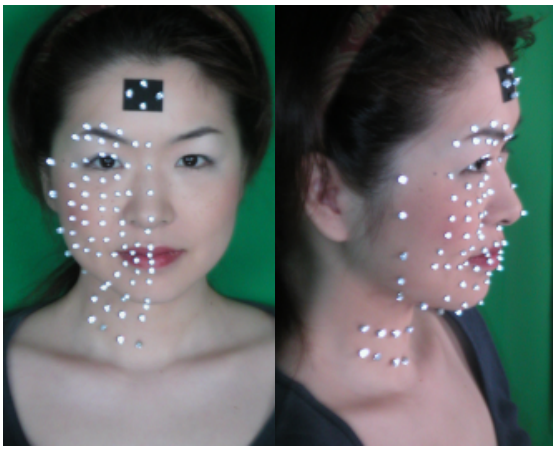


図4　マーカの配置：顔半分

3.2. データの取り込み

モーションキャプチャデータは6台の赤外線カメラ画像からそれぞれのマーカの位置を求めていく。その際、データの算出ソフトウェアであるVicon Workstation、Vicon IQと呼ばれる専用アプリケーションを用いて、データをC3Dフォーマット[5]にコンバートした。C3Dフォーマットは3次元のバイオメカニクスデータを管理する標準フォーマットで、臨床歩行、バイオメカニクス研究やモーションキャプチャスタジオ等で広く使用されている形式である。制作したモーションキャプチャ（C3D）データのフレームレートは120[Hz]であり、137点（顔半分の場合78点）の毎フレームの位置が格納されている。モーションキャプチャシステムとDVCAMコーダのタイムコードが同期しているため容易に発話した文章、単語単位にファイル化することができる。

モーションキャプチャデータの切り出し、さらに次節で述べるキャプチャーデータのデータ編集を容易に行う必要があるため、3D モデリング、アニメーション、レンダリング用ソフトである Alias 社の Maya 5.0 上にデータの読み込み、書き出しができるように Plug-in 開発用ライブラリである Maya API を用いて独自にデータ入出力 Plug-in を制作した。図 5 に読み込み後のマーカデータを示す。

3.3. データの正規化

獲得した各マーカのデータは対応する顔部位の変化量と発話時の頭部の動きが含まれており、コーパス作成時は実際のマーカの変化量のみを取り出す（正規化する）必要がある。従って頭部の動きのみが含まれていると考えられる補正用マーカ、額部：4 点、こめかみ部：2 点および鼻部：2 点を用いて頭部の回転を求める。 頭部変化前のマーカの 3 次元データを同次座標系で $\mathbf{P} = \langle P_x, P_y, P_z, 1 \rangle$ として表現し , 変化後の点を $\mathbf{P'} = \langle P'_x, P'_y, P'_z, 1 \rangle$ とするとアフィン行列 $\mathbf{M}$ は次のように簡単に表すことができる。

$$\mathbf{P'} = \mathbf{MP} \quad \mathbf{M} = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$\mathbf{M}$ を補正用マーカ 8 点から特異値分解により求めることができる。但し、首部のマーカは頭部の動きに影響を受けず、首自身の動きに影響を受けるため、首の正規化は別途、補正用マーカ 4 点を用意し同様の処理を行った。

4. モーションキャプチャデータから顔モデルへの対応付け

本章では正規化後のモーションキャプチャデータから顔モデルへの対応付けの方法を述べる。顔モデルは擬人化エージェント Toolkit:Galatea[6]のエージェント生成・表示部 Face Synthesis Module（FSM)[7]で使用されている顔モデルを採用した。FSM では正面顔写真と

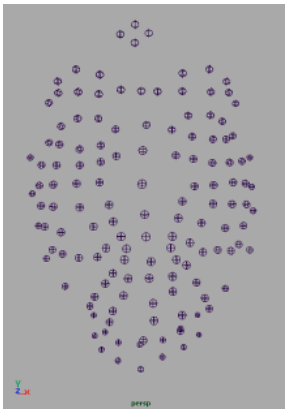


図 5　マーカデータの読み込み

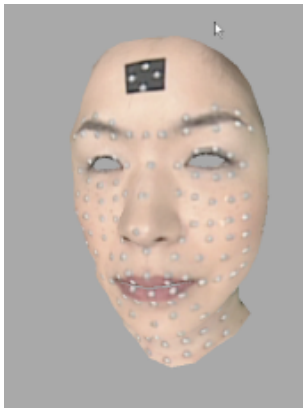


図 6　整合後の顔モデル

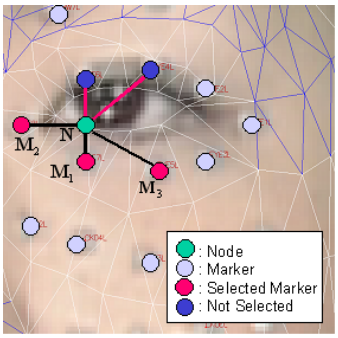


図 7　近接したマーカの選択法

モジュール内に用意されているワイヤフレームモデルを整合するツールが含まれている。今回このツールを使用し発話者の正面画像を用いて顔モデルを整合した。整合後の顔モデルを図 6 に示す。このモデルは三角形ポリゴンで約 750 ポリゴンで構成されている。またこれらポリゴンは頂点とエッジを他のポリゴンと共有化されている。顔整合ツールではマニュアルにて整合を行うが、20 点程度の指定された特徴点のみを動かせば、適切なモデルが完成する。整合は 2 次元の画像で行うため、モデルの奥行き方向の整合は行われない。さらにこのときの各マーカの 2 次元座標を獲得する。

顔モデルのアニメーションはモデルの各ノード（頂点）にマーカの変化量を割り当てることで実現させる。割り当て方法は図 7 のようにモデルのノードから最も近い数点のマーカを選択し、マーカとノードの距離に応じてマーカが持つ変化量をノードに割り当てる。あるノードの座標を $\mathbf{N}$ とし、その周囲にあるマーカの座標を $\mathbf{M}_i$、表情変化後のマーカ座標を $\mathbf{M'}_i$ すると変化量 $\mathbf{v}$ は次の式で与える。

$$\mathbf{v} = \sum_{i}^{n} (\mathbf{M'_i} - \mathbf{M_i}) \cdot \left(\frac{1.0}{\mathbf{N} - \mathbf{M_i}}\right) \cdot \left(\sum_{i}^{n} \frac{1.0}{\mathbf{N} - \mathbf{M_i}}\right)^{-1}$$

n は選択したマーカ数を示し、n=3 とした。しかし、下唇付近に存在するノードに移動量を割り当てる際、上唇付近のマーカを選択してしまう場合がある。そのため適切な値を求めることが発生し、同様な問題が口や目の境界エッジ（共有されていないエッジ）部分で起こることがある。そのためマーカの選択時にマーカとノードを結んだ線分が顔モデルに存在するすべての境界エッジを跨ぐか判定し、跨いだ場合そのマーカは選択対象から外すことで問題を解決している。割り当て後の顔形状変化後のモデルを図 8 に示す。

5. おわりに

音声と動画像、そしてモーションキャプチャによる顔部位の変化量を含むマルチモーダルコーパスを構築した。このコーパスは ATR 日本語バランス音律文を発話用テキストとし、日本人女性 1 名によって構成されている。さらにキャプチャしたモーションキャプチャデータから顔モデルへの割り当て方法について述べた。これにより、大量のマーカで撮影した場合でも容易に顔モデルのアニメーションが実現できる。今後の展開として、発話サンプル数を増やし、さらに日本語のみならず英語発話によるコーパス収集が挙げられる。

謝辞：本研究は情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

参考文献

[1] Nakamura, S. ”Overview on Recent Activities in Multi-Modal Corpora”, COCOSDA Workshop, 2000.

[2] Ezzat, T., Geiger,G. and Poggio,T.“ Transable Videorealistic Speech Animation ”, Proceedings of ACM SIGGRAPH 2002, 2002.

[3] Morishima, S., Iwasawa, S., Sakaguchi, T., Kawakami, F., and Ando, M.,“ Better Face Communication ”, Visual Proceedings of ACM SIGGRAPH'95, Interactive Communities, p.117, 1995.

[4] Tamura, T., Kondo, S., Takashi Masuko, T. and Kobayashi, T.“ Text-to-Audio-Visual Speech Synthesis Based on Parameter Generation from HMM ”, Proceeding of EUROSPEECH, pp.959-962, 1999.

[5] The C3D web site [http://c3d.org]

[6]Kawamoto,S., Shimodaira, H., Nitta, T., Nishimoto, T., Nakamura, S., Itou, K., Morishima, S., Yotsukura, T, Kai, A., Lee, A., Yamashita, Y., Kobayashi, T, Tokuda, K., Hirose, K., Minematsu, N., Yamada, A., Den, Y., Utsuro, T., and Sagayama, S., "Open-source software for developing anthropomorphic spoken dialog agent" Proceedings of PRICAI-02, International Workshop on Lifelike Animated Agents, pp.64-69, 2002

[7] Yotsuskura, T., Morishima, S., and Nakamura, S., "Model-based Talking Face Synthesis for Anthropomorphic Spoken Dialog Agent System", ACM Multimedia Conference 2003, Proceedings of the 11th ACM International Conference on Multimedia pp.351-354, 2003

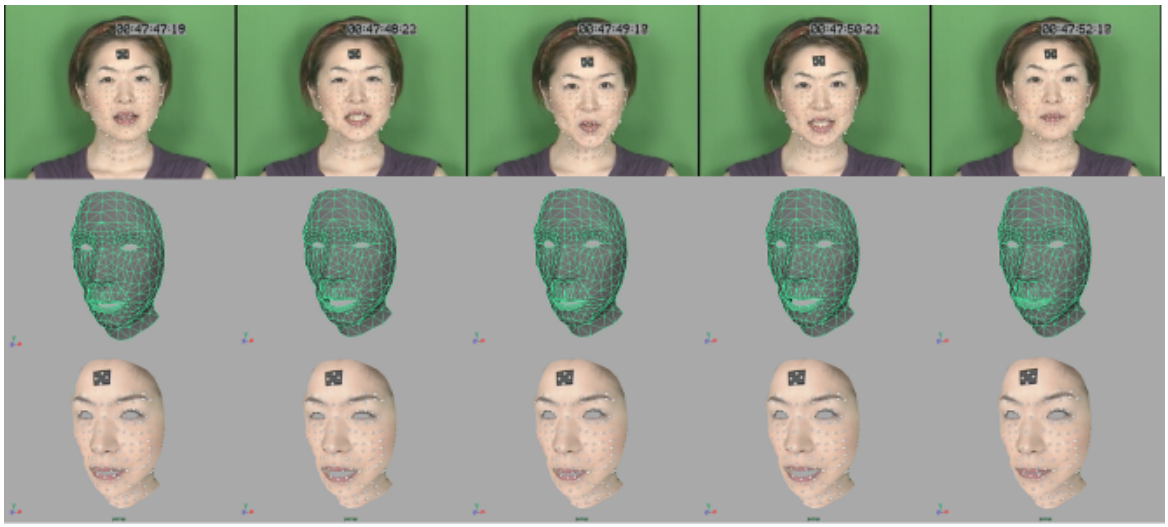


図 8　モーションキャプチャデータを用いた発話アニメーション