情報処理学会論文誌 Vol.56 No.8 1680–1690 (Aug. 2015)

HMMによる電子掲示板からのトークアニメの生成

加藤 和輝[1] 今村 友哉[1] 石本 貴康[1] 渡部 郁美[1] 山本 正信[1,a)]

受付日 2014年10月23日, 採録日 2015年5月9日

概要：本論文では，シナリオの台詞から直接アニメを制作する手法を提案する．まず，ビデオ映像から人物の発話と身振りを抽出し，発話を出力，身振りを状態とし，両者の関連 HMM（隠れマルコフモデル）で学習する．この関連を使って，台詞から HMM により状態遷移，すなわち身振りを生成する．このとき，台詞の単語を上位概念で置き換えることにより，少ない学習データからでも身振りが生成できることを示す．本手法によるアニメの制作例として，電子掲示板の書き込みからトークアニメを作成する．

キーワード：キャラクタアニメーション，隠れマルコフモデル，電子掲示板，会話

HMM-based Talking Avatars from Electronic Bulletin Board

Kazuki Kato[1] Tomoya Imamura[1] Takayoshi Ishimoto[1] Ikumi Watanabe[1]
Masanobu Yamamoto[1,a)]

Received: October 23, 2014, Accepted: May 9, 2015

Abstract: This paper proposes a method to translate words in a scenario to an animation movie directly. First, a motion capture system captures a series of poses of a person with speech from video clips. The parameters of the HMM (Hidden Markov Model) are estimated from the speech (output) and pose (state). The HMM decodes words in a scenario as the state transition, i.e. a gesture. At this time, substituting some words with a dominant conception, the gesture can be generated even from small learning set. It is shown that the proposed method produces talking avatars from writes on the electronic bulletin board.

Keywords: character animation, HMM, electric bulletin board, dialogue

1. はじめに

Web の電子掲示板では，特定の話題について意見を書き込むと，他の人がそれに対して意見を書き込み，さらに別の人が意見を書き込む．このような一連の書き込みは議論といってよい．アニメキャラクタに書き込みを発話させれば，テレビのトーク番組や討論番組のようなコンテンツとなるのではないだろうか．たとえば，「2 チャンネル」での 2004 年 3 月のある書き込みは，「電車男」として単行本化され，トーク番組とは異なるが，後にテレビドラマや映画として映像化された．書き込みの映像化は，特定のグループだけではなく一般の人々の耳目を集め，議論の共有化と深化が期待できる．

本研究の目的は，掲示板ログを議論として再構成したトークアニメを作成することである．これまでのアニメの制作では，先にアニメ映像を作成しアフレコで台詞を吹き込んできた．本研究では，掲示板ログを直接アニメに変換する手法の開発を目指している．これは，アニメ制作の専門的スキルを持たないユーザクリエータのためのアニメ制作法として期待できる．

掲示板ログの書き込みを発話させたとき，発話と直接関連するのは口であり顔の表情である．発話に連動した顔の表情生成に関する研究はすでにある [2], [3]．本研究では身振りに着目し，実際の対話映像から発話と身振りの関係を HMM（隠れマルコフモデル）で学習させ，台詞が与えられたとき，発話にふさわしい身振りを生成して HMM でアニメ映像を作成する手法を提案する．

[1] 新潟大学
Niigata University, Nishi, Niigata 950–2181, Japan
[a)] yamamoto@ie.niigata-u.ac.jp

2. 関連研究

発話と身振りが互いに関連があることは知られていた．心理学者の McNeill [9] は，身振りをその発話と関連付けて，iconic，metaphoric，deictic および beat の 4 種類のジェスチャーに分類した．

発話に身振りが対応付いているならば，シナリオの台詞に身振りを注釈として加えることができる．Cassell ら [4] は構文解析により，Neff ら [10] は統計的手法により台詞の分節や語彙に身振りの添付を行った．さらに Marsella ら [8] は，台詞を発話した音声から感情を抽出し身振りに反映させた．また，Sun ら [12] は対話の形式に応じてシナリオの選択を提案した．台詞を俳優に発話させあるいは音声合成させ，注釈に基づきキャラクタを動作させればアニメーションが得られる．これらは，シナリオの拡張という意味でスクリプトベースの発話アニメーションと呼ぶことにする．

一方，モーションキャプチャが汎用的な技術となり，スピーチ付の身振りデータが容易に得られるようになった．スクリプトベース手法は，演者の観察に基づき，身振りの注釈を手動で加えている．これに対し，モーションキャプチャの利用は，発話と身振りの関係が機械的に学習できるようになった．

顔の表情，特に口の開閉は発話と直接関連している．スピーチの音韻と映像の関係を，Bregler ら [3] は N グラムモデルにより，Brand [2] は HMM により学習した．それぞれ，新たなスピーチに対応する顔の映像生成が可能である．

スピーチと身振りの対応付けに基づくアニメーションは，Stone ら [11] により提案された．Stone らは台詞を句構造に構文解析し，句と身振りの対応付けから，句の組合せによりアニメーションを生成した．Levine ら [7] は，句をさらに細かい音節に分解し，音節と身振りの関連を HMM により学習した．音節は句よりも細かな単位であるので，句の組合せよりも音節の組合せのほうがアニメ化できる台詞が広がるとされる．これら，スピーチと動作の事例からの学習に基づく手法を事例ベースの発話アニメーションと呼ぶことにする．

事例ベース法はスクリプトベース法に比べ，より少ない手動介入でアニメーションを作成することができる．しかし，どのようなスピーチに対してもアニメーションを作成するためには，膨大な学習データを必要とする．

3. 提案手法の概要

映画やドラマの制作はもとより，アニメの制作でもキャラクタの身振りには，視聴者をひきつける要素が必要である．Stone ら [11] や Levine ら [7] は，プロの俳優に依頼し学習用演技の収録をスタジオで行っている．トークアニメを制作するには，実際のトーク番組が参考になる．そこでは，多数の芸能人が出演している．そこで，テレビ番組から芸能人の漫才を収録し，その発話と身振りの関係を学習することにする．

Levine ら [7] は，俳優の音声と身振りとの関係を HMM で学習したが，テレビ番組から収録した映像音声には BGM や笑い声のようなノイズが多く含まれている．本研究では，映像音声に含まれる発話を記号化し，発話文字列と身振りを HMM で学習する．

発話を文字列として表すことにより，類語や語彙の上位概念を利用して学習データの貧弱さを補填することが可能となる．たとえば，「わたし」と発音されたスピーチによる学習では，「ぼく」と発音された入力に対する動作はない．しかし，「ぼく」も「わたし」も「自分」という上位概念の例としておけば，「ぼく」の動作は「わたし」の動作で置き換えればよい．提案手法は，事例ベースとスクリプトベースを融合させた手法ともいえる．

4. 発話と身振りの関連付け

実際に人が演じているテレビ映像から演者の発話と身振りの関係を HMM により学習する．学習データとして，2007 年に放映されていたテレビ映像から，31 本のビデオクリップを収録した．ビデオクリップのフレーム総数は 1,935 フレーム，約 1 分強であった．このビデオクリップには，12 組の芸能人の漫才が収められている．ビデオクリップは，身振り映像ファイルと音声ファイルに分離される．学習の準備として，映像からモーションキャプチャ [15] による演者の身振りの測定と記号化，および映像音声に含まれる発話の記号化を行っておく．身振りを状態，発話を出力として HMM のパラメータ推定を行う．付録 A.1 には，ビデオクリップのファイル番号，発話者によるビデオクリップ名，フレーム数，発話内容が順に示されている．

4.1 身体多関節モデル

本論文で使用する身体の多関節モデルは，腕や脚などの 16 個の部位が関節で結合したリンク構造である．**図 1** はこの構造の多関節モデルである．このリンク構造は，最上位の部位を腰部とする階層構造である．連結した部位のう

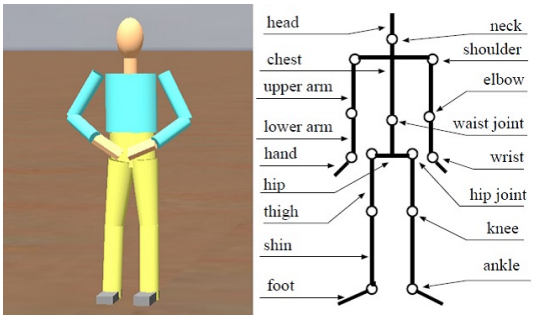


図 1 身体多関節モデル
Fig. 1 An articulated model of human body.

ち腰部に近い部位を親部位，遠い部位を子部位とする．それぞれの部位は固有の座標系を持ち，部位の姿勢は親部位の座標系に対する姿勢で表す．ここで，姿勢パラメータはロール・ピッチ・ヨー角で表す．最上位の部位である腰部の親部位はカメラとする．腰部の並進運動を除くすべての部位の姿勢パラメータの集まりをモデルの姿勢とする．姿勢は 48 $(= 16 \times 3)$ 次元ベクトルである．

4.2 身振りの記号化

姿勢は実数ベクトルであるが，HMM の状態として扱うために記号化しておく．記号化はクラスタリングにより行い，各クラスの代表姿勢がコードワードでありコードブックを構成する．学習姿勢列は代表姿勢列に記号化される．これを学習姿勢コード（略して学習コード）と呼ぶ．

クラスタリングは，k-means 法 [14] を使用した．学習姿勢ベクトル集合に対し，与えられたクラス数の初期代表姿勢を用意し，代表姿勢とのユークリッド距離により姿勢ベクトルをクラス分けする．代表姿勢をクラス内の平均ベクトルとして更新する．この更新とクラス分けをクラス分けの変更がなくなるまで繰り返す．本論文の実験では，各ビデオクリップに対し，20 フレームおきにピックアップした姿勢に正立姿勢を加えた 109 姿勢を初期代表姿勢とした．すなわち，$k = 109$ として，全 1,935 フレームの学習姿勢列に対しクラスタリングを行った．その結果得られた 109 個の代表姿勢を**図 2** に示す．

連続動作からのクラスタリングであるので，得られた学習コードには，同じ代表姿勢が並ぶことが多い．学習コード上で連続して現れる代表姿勢の列を学習姿勢セグメント（略して学習セグメント）と呼ぶ．学習コードの一部を**表 1** に示す．第 1 列は，全ビデオクリップ上でのフレームの通し番号，第 2，3 列は，それぞれ付録 A.1 のビデオクリップファイル番号とそのファイル上でのフレーム番号，第 4 列が代表姿勢コードである．代表姿勢 76，3，77 の学習セグメントが見られる．

図 3 は各学習セグメントのフレーム長を 3 次元棒グラフで示している．定義域の横軸がクラスタ番号，縦軸が学習セグメントの個数である．クラスごとにセグメント長の順番に異なる色で色付けされて配列されている．学習セグメント長の平均は 8，最長は 52，最短は 1 フレームであった．また，各姿勢クラスに含まれる学習セグメントの個数は，最大 10，最小 1，平均 2.2 個であった．

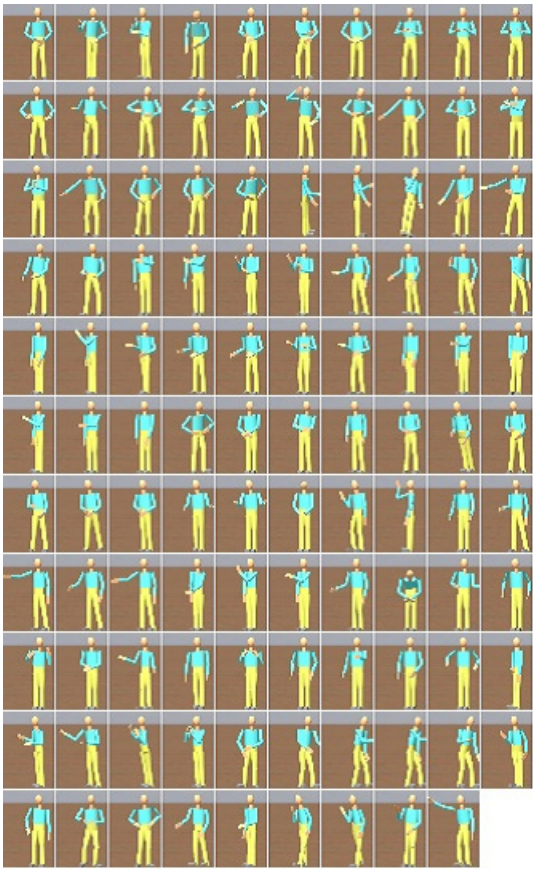


図 2 代表姿勢
Fig. 2 Poses as code words.

表 1 姿勢データのクラスタリング
Table 1 Clustering of poses.

G. frame	File	L. frame	Pose code	Letter
.	.	.	.	.
1367	22	002	76	お
1368	22	003	76	お
1369	22	004	3	ん
1370	22	005	3	ん
1371	22	006	3	な
1372	22	007	3	な
1373	22	008	77	な
1374	22	009	77	で
1375	22	010	77	で
1376	22	011	77	で
1377	22	012	77	す
1378	22	013	77	す
.	.	.	.	.

4.3 発話の記号化

身振り映像と同期して発せられる音声を，HMM の出力として扱うために記号化しておく．テレビ映像の音声から演者の発話を抽出し文字列に留める．映像音声には，演者の発声以外にも相手方演者の発声や観客の笑い声，BGM なども含まれている．そのため，映像音声から自動的に演者の発声のみを文字列に変換するには困難さをともなう．ここでは，音声を聞きその wave 信号を見ながら手動で変換することにする．なお，ここでの文字列とは，発声した音声を文字に変換したときの文字の時系列である．たとえば，「ひま（暇）」の発音は，「ひひひひままま」のように文字列化されることもある．これは，時間単位をビデオ映像

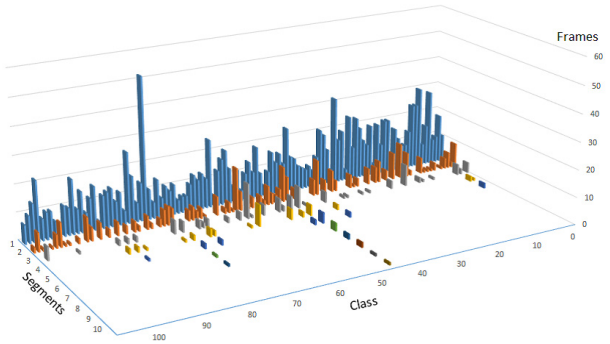


図 3 学習姿勢セグメント
Fig. 3 Pose segment.

の 1 フレーム（1/30 秒）としているため，1 秒間に 30 回のサンプリングでは同じ文字が連続して現れる可能性があるためである．この文字列を発話文字列と呼び，書き言葉の文字列とは区別する．発話文字列は時間情報を含んだ文字列である．表 1 の第 5 列に，発話文字列の例が示されている．

4.4 HMM による学習

フレーム時刻での発話と姿勢が得られたとする．HMM の状態を姿勢に，出力記号を発話文字に対応させて，HMM のパラメータ，すなわち，状態遷移確率，出力記号確率および初期状態確率を推定する．通常，これらのパラメータは，出力記号列からバウム・ウェルチアルゴリズムによって推定される．しかし，HMM では隠されている状態が，本研究では姿勢として観測可能であるため，状態遷移確率，出力記号確率は，代表姿勢や発話文字の出現頻度から直接求めることができる．初期状態確率は，使用したビデオクリップからその先頭姿勢の出現頻度より求める．

5. 対話のアニメ化

掲示板ログの書き込みから，身振りを交えて発話するキャラクタのアニメ映像を HMM により制作する．まず，掲示板ログの書き込みを，議論の様式に沿うように整理し，対話のシナリオを作成する．台詞を時間軸が含まれた発話文字列に変換する．HMM により，台詞の発話文字列を出力記号としたときの，最適な状態（姿勢）遷移，すなわち身振りを求める．

5.1 掲示板ログの対話シナリオ化

掲示板ログの書き込みを対話のシナリオに変換する．現在，この変換は手作業による編集である．

5.1.1 ログ編集

掲示板の書き込みは，さまざまな話題が同時に進行している．書き込まれた順番そのままにアニメーション化を行っても会話の意味が通じない．したがって書き込みの順番を入れ替えて，会話の流れを編集する必要がある．

書き込みには，アスキーアート，URL，顔文字といった音声にできないものが含まれる．また，話題に無関係な書き込みもある．そのようなものを削除する必要がある．大まかに分けて，次の 2 段階の編集作業が必要である．

5.1.1.1 個々の書き込みに対する編集

- 明らかな誤字や脱字などのミスの修正：電子掲示板独特の書き方として，故意に誤字や誤読をしている場合もある．
- 話題にまったく関係ない無意味な書き込みの削除：荒らし行為や宣伝，マルチポストによる書き込みなど．
- URL，顔文字，アスキーアートの削除：これに関連し，URL リンク先の感想の書き込みなども，削除する必要がある．
- 返信先のミス，掲示板ごとの返信ルールにそぐわない部分の修正：たとえば後々に，「先ほどの書き込みの返信相手が違った」という書き込み．

5.1.1.2 ログ全体に対する編集

全体を編集し，掲示板のログを複数人による会話の形式にして，アニメ化しやすくする．

- 誰に対する返信かという情報を利用した順番の入れ替え：基本は掲示順である．ある書き込みに注目したとき，その書き込みに対する返信が見つかった場合，その書き込みを直後に持ってくる．もし返信をたどった際に，まだ出ていない書き込みを参照している場合は，直前に持ってくる．
- ハンドルネーム，ID，IP アドレスなどの情報を使い，同一の発言者による書き込みの特定：キャラクタの人数を決定するため，発言者の特定を行う．ID，IP アドレスが表示されない掲示板も多い．ID は日付によって，IP アドレスは接続時ごとに変わるので，掲示板全体を通して完全に個人を特定するには情報不足である．書き込みに，変わる前の ID や前に書き込んだ際のレス番が載っている場合もあるので，書き込み内容にも注目する必要がある．ハンドルネームだけによる判断では，以前に書き込んだ誰かを騙る書き込みがあった場合に問題があるため，注意が必要である．

これらの手順だけで完全に自然な会話の流れになるのかといえば，必ずしもそうではない．編集された文章を読み，さらに自然な会話になるように再編集する必要が出てくることもある．

5.1.2 対話シナリオ

実際の掲示板ログの一部を付録 A.2 に示す．このログは，「まち BBS　北陸・甲信越 新潟県新潟市スレ【79】226〜231」に掲載されている．このログを編集して以下の対話シナリオを作成した．キャラクタ名，スレ番号，発話内容の順に記す．

A 226　何度か書き込みしてますが　新潟市内で海鮮メインで夜遅くまで営業してる居酒屋教えてください　29

日火曜に行きます

B 227 海鮮居酒屋はなの舞はどう 駅前だし

A 229 はなの舞ってチェーン店居酒屋ですよね せっかくの情報ですがわざわざ新潟まで行くので有名オススメ居酒屋を知りたいのです 私の聞き方悪くてごめんなさい

C 228 壱勢とか

D 231 遅くまでって何時頃 新潟は早じまいが多いよ海鮮だと大助はお勧めだけど 23:00 あと葱ぼうず 0 時だったかな 新潟のグループでは安兵衛 船栄 ちょこざいや系とかがメジャー 後はホットペッパー検索しておすすめレポートなど参考にした方が吉 店の雰囲気も分るしお勧め

5.2 対話の音声化と記号化

シナリオの台詞には発話の時間情報が含まれていない．HMM からの身振りの生成には，台詞を時間軸上の発話文字列に変換する必要がある．そのためには，台詞を発話により音声信号化すれば，3.2 節で述べた方法により手動で発話文字に変換可能である．本論文では，音声合成ソフトにより音声信号に変換した．音声合成ソフトは発話の速度などが制御可能なので，台詞を発話文字列へ自動変換することが，原理的には可能と考えられる．しかし，使用した音声合成ソフトは合成過程がブラックボックス化されているため自動化が難しく，面倒ではあるが合成音声を手動で発話文字列へ変換した．

音声合成ソフトは，当初有償の VoiceTEXT [13] を使用していたが，Windows XP から 7 への切替えにともない，フリーソフトの CeVIO [5] に切り替えた．両者とも，

- 比較的自然な音声を合成できる，
- wave ファイルでの音声出力ができる，
- 話すスピード，声の高さが設定できる，
- 感情を表すバロメータが存在し，セリフに感情を乗せることができる，

といった機能がある．

VoiceTEXT には男性 2 名と女性 2 名の音源が備わり，CeVIO には女性 1 名の音源が備わっている．後者は，音源の声質や高さを変えて，登場する 4 名の音声を区別させた．

5.3 HMM による生成

HMM により，発話記号列を出力とする最適な姿勢列を求める．これは HMM における復号問題であり，効率的なアルゴリズムとしてビタビ・アルゴリズムが知られている．このとき，復号の中断を避けるために，ゼロ頻度問題を解決しておく必要がある．復号された姿勢列を生成姿勢コード（略して生成コード）と呼ぶ．アニメキャラクタへは，この生成コードを姿勢パラメータに変換して与えることになる．しかし，生成コードには同じ代表姿勢が連続して現れることが多く，単調な動作となる．この生成コードを動きのある身振りに変換する．

5.3.1 ディスカウンティング

HMM 内の状態遷移などの確率分布は，本論文では発話と身振りの学習データから計算されている．しかし，学習データには限りがあるため，本来遷移すべき状態間の遷移確率がゼロになるというゼロ頻度問題が起きる．そのため，状態が遷移せず身振りの生成が中断する．

まず，各状態から全状態への遷移を加えることも考えられる．しかし，これは学習データに存在しない遷移が発生し，学習結果が生成した身振りに反映されない可能性があるので，本研究では採用しない．

そこで，同じ状態が継続する可能性は考えられるので，各状態から自身への遷移がなければそれを加え，すべての出力記号が各状態から出力可能とする．具体的には，各状態から自状態への遷移頻度に 1 回，各状態から全出力記号の出力頻度に 1 回ずつ加えて，状態遷移確率分布，記号出力確率分布を再計算する．これにより，どのような発話文字列からでも必ず代表姿勢列が生成されるようになる．

このような修正は身振りの生成が止まることはないが，同じ姿勢が連続して現れる恐れがある．動きのある身振りを生成するための方法を次節で示す．

なお，Stone ら [11] は，句と動作の対からデータベースを構成し，台詞中の句とデータベース中の句の照合により身振りを生成している．照合できなければ身振りの生成が止まることになる．

5.3.2 生成コード列から動作への復号

自状態への遷移の追加を行ったことに加え，代表姿勢は一定の変動幅内の姿勢を代表している．そのため，HMM の復号により得られる生成コード列には同じ代表姿勢が連続して現れることが多くなる．これは，一定の姿勢が継続される一方，代表姿勢の変わり目には姿勢のギャップを生じやすく，ぎこちない動作が生成される．動作が自然で滑らかになるように生成コードから動作への変換法を示す．

学習姿勢列はクラスタリングにより学習コードに変換されるのであった．このとき，学習姿勢列の姿勢パラメータは，学習コードの代表姿勢とフレーム順に一対一に対応している．そこで，生成コードの代表姿勢列が，学習コードの代表姿勢列と一致しているならば，学習コード列に対応する姿勢パラメータ列を生成動作とする．このようにすれば，自然で滑らかな身振りが得られる．これは，姿勢パラメータから代表姿勢へのコード化の逆であるので，復号化と呼んでもよい．

問題は，生成コードの代表姿勢列と一致する学習コードの代表姿勢列が複数存在する場合である．学習コード内で連続する代表姿勢列を学習セグメントと呼んでいた．同様に，生成コード内で連続する代表姿勢列を生成セグメントと呼ぶことにする．学習コードも生成コードも複数のセグ

メントの連結により構成されている．

生成コードをそのセグメント順に復号する．このとき，先頭セグメントの復号と復号の終わった生成セグメントに続く中間セグメントの復号とでは復号方法が異なる．

(1) 生成セグメントが先頭セグメントであるとき，生成セグメントに長さが最も近い学習セグメントを選ぶ．それらが複数存在する場合はランダムに 1 つ選ぶ．生成セグメントを選ばれた学習セグメントで，(3) の方法で復号する．

(2) 生成セグメントが中間セグメントであるとき，復号が終わったばかりの生成セグメントと中間セグメントから構成されるセグメント対を考える．このセグメント対と同じ代表姿勢を持つ学習セグメント対が学習コード内に存在するならば，中間セグメントに対応する学習セグメントで中間セグメントを (3) の方法で復号する．ただし，対応する学習セグメントが複数個存在するならば，中間セグメントに長さが最も近い学習セグメントを選ぶ．学習セグメント対が見つからない場合には，中間セグメントを先頭セグメントと見なし，(1) の方法で対応する学習セグメントを見つける．

(3) 生成セグメントの長さが，学習セグメントよりも短い場合は，生成セグメントを学習セグメントに対応する姿勢で先頭から順次復号する．生成セグメントの長さが，学習セグメントよりも長い場合は，学習セグメントを引き延ばして，生成セグメントに当てはめる．具体的には，生成セグメントの長さを学習セグメントの長さで割り，その商と余りを求める．商の数だけ学習セグメントの姿勢を繰り返す．余りがあれば，先頭姿勢から順に余りの数だけ，姿勢の繰返し数を 1 つ増やす．

修正後の生成動作は，セグメント間の動作のつなぎ目が不自然になることもあるので，カルマンスムーザにより動作を滑らかにする．

5.3.3 動作の生成実験

実際に，書き込み 227 番の発話「海鮮居酒屋　はなの舞はどう　駅前だし」から動作の生成過程を示す．発話していない時刻も含まれるが，無音時の動作も学習しているので，無音でも動作の生成が可能である．

5.3.3.1 ディスカウンティングを行わない場合

生成コードを**表 2** の生成コード 1（Pose code 1）の列に示す．このときの HMM のトレリスを**図 4** に示す．最上行が時刻（フレーム），その下が状態（代表姿勢）の遷移，最下行が出力の発話文字で，この文字を出力できる代表姿勢番号が楕円で囲まれている．

31 本のビデオクリップの先頭代表姿勢から初期確率を得る．ほとんどの代表姿勢がたかだか 1 回だけ現れるが，代表姿勢 “47” のみが 2 回現れ初期確率最大となった．また，すべての代表姿勢から無音が出力されている．結局，初期確率と出力確率の積が最大となった代表姿勢 “47” が，先頭フレームの姿勢となり，次のフレームでも継続する．第 1 フレームの姿勢 “47” から遷移可能な代表姿勢は “40”，

表 2 書き込み 227 からの動作生成
Table 2 Gesture from writting # 227.

Frame	Letter	Pose code 1	Pose code 2	Pose frame	Frame	Letter	Pose code 1	Pose code 2	Pose frame
0		47	47	12:00	46	ま		40	12:44
1		47	47	12:01	47	い		52	15:39
2	か	89	89	19:36	48	い		52	15:40
3	か	89	89	19:36	49	い		52	15:41
4	か	89	89	19:36	50	い		52	15:42
5	い	89	89	19:37	51	は		52	15:43
6	せ		89	19:37	52	は		52	15:44
7	せ		89	19:37	53	は		52	15:45
8	せ		89	19:38	54	は		52	15:46
9	せ		89	19:38	55	ど		52	15:47
10	ん		89	19:39	56	ど		52	15:48
11	ん		89	19:39	57	ど		52	15:49
12	い		89	19:40	58	う		52	15:50
13	い		89	19:40	59	う		52	15:51
14	い		89	19:41	60	う		52	15:52
15	い		89	19:41	61	う		52	15:53
16	ざ		89	19:42	62	う		52	15:54
17	ざ		89	19:42	63			47	19:49
18	ざ		89	19:43	·	·	·	·	·
19	ざ		89	19:43	75			47	20:03
20	ざ		89	19:44	76	え		89	19:36
21	か		89	19:44	77	え		89	19:37
22	か		89	19:45	78	え		89	19:38
23	や		89	19:45	79	え		89	19:39
24	や		89	19:46	80	き		89	19:40
25	や		89	19:46	81	き		89	19:41
26	や		89	19:47	82	き		76	19:09
27	や		89	19:47	83	ま		3	5:57
28	や		89	19:48	84	ま		3	5:58
29	や		89	19:48	85	え		4	2:17
30			47	19:49	86	え		4	2:18
·	·	·	·	·	87	え		4	2:19
35			47	19:54	88	だ		4	2:20
36	は		40	12:37	89	だ		4	2:21
37	は		40	12:37	90	だ		4	2:22
38	は		40	12:38	91	し		5	2:45
39	な		40	12:38	92	し		5	2:46
40	な		40	12:39	93	し		5	2:47
41	の		40	12:39	94	し		5	2:48
42	の		40	12:40	95	し		5	2:49
43	ま		40	12:41	96	し		5	2:50
44	ま		40	12:42	97	し		4	2:60
45	ま		40	12:43	98			0	2:61

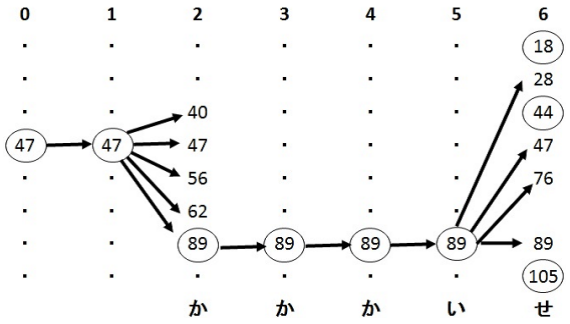


図 4 トレリス
Fig. 4 Pose code word.

“47”，“56”，“62”，“89” の 5 つであるが，“89” のみが「か」を出力可能であるので，第 2 フレームの姿勢は “89” となった．この後，自状態への遷移確率が最も高かったため，自状態が継続している．

5 フレーム目の代表姿勢 “89” は，代表姿勢 “28”，“47”，“76”，“89” に遷移可能であるが，遷移先のいずれも「せ」を出力することができない．そのため，5 フレーム目までしか動作を生成することができなかった．

5.3.3.2 ディスカウンティングを行った場合

どの状態からでも，あらゆる発話文字が出力できるようにした．これにより，5 フレーム目の状態 “89” から自状態へ遷移し，発話文字「せ」を出力させることができた．得られた生成コードを表 2 の生成コード 2（Pose code 2）の列に示す．遷移の停止はなくなったが，自状態への遷移が多く，代表姿勢列による生成動作は単調になる．

5.3.3.3 生成コードの復号

HMM からの生成コード列を姿勢パラメータ列に復号する．復号結果を表 2 の姿勢フレーム（Pose frame）の列に示す．コロンの左がビデオクリップのファイル番号，右がそのファイル上のフレーム番号である．このフレーム番号の画像から得られる姿勢パラメータに復号するのである．

まず，代表姿勢 “47” から構成される先頭生成セグメントを復号する．学習コードから，代表姿勢 “47” の学習セグメントを図 3 から抜き出したのが**図 5** (a) である．全部で 10 本の学習セグメントのうち，生成セグメントの長さ 2 と同じ長さの 7 番の学習セグメントが選ばれた．この学習セグメントの学習コード上のビデオクリップ番号とフレーム番号が姿勢フレームの列に記されている．

次の生成セグメントは，代表姿勢 “89” から構成されている．長さ 2 の代表姿勢 “47” の学習セグメントに続く代表姿勢 “89” の学習セグメントは，学習コード上に存在しなかったので，この生成セグメントは先頭セグメントとして復号する．この先頭セグメントの長さは 28 フレームある．一方，代表姿勢 “89” の学習セグメントを図 3 から抜き出したのが図 5 (b) である．5 本の学習姿勢列のうち最長 13 フレームの学習セグメントでも生成セグメントの長さに及ばない．そこで，学習セグメントを伸ばして生成セグメントに対応付ける．$28 \div 13 = 2$ あまり 2 であるので，学習セグメントの最初の 2 フレームの姿勢をそれぞれ 3 回コピーし，残りは 2 回ずつコピーする．

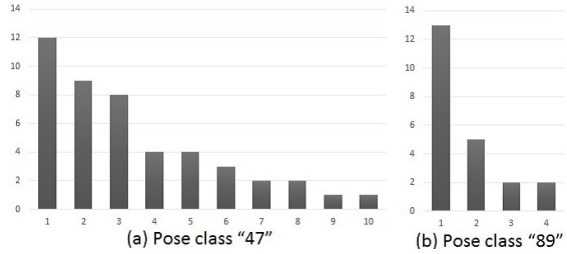


図 5 学習セグメントの長さ．(a) 代表姿勢 “47”，(b) 代表姿勢 “89”
Fig. 5 Length of segments. (a) pose class “47”, (b) pose class“89”.

第 30 フレームからは再び無音が続き，代表姿勢は “89” から “47” に切り替わる．このとき，代表姿勢 “89” の学習セグメントは，代表姿勢 “47” の学習セグメントに学習コード上でつながっている．そのため，学習コード上のフレーム番号で連続して置き換わっている．

書き込み 227 番の書き言葉の文字列のうち，付録 A.1 に示される学習データと照合できる文字列の長さはたかだか 2 であった．本手法は，このような断片的な照合からでも身振りが生成できている．その理由を考察する．

図 4 での代表姿勢 “47” から “89” への遷移は，ビデオクリップのファイル 8 と 19 に表れているが，ともに代表姿勢 “89” から「か」を出力していない．「か」は，ファイル 26 で代表姿勢 “89” から自己遷移したとき出力している．記号出力確率が状態遷移に依存していないため，実例にない遷移が可能となる．また，ディスカウンティングも，身振りの生成を継続するためには必要ではあるが，実例にはない遷移を作り出している．さらに，生成セグメントを，発話文字ではなくセグメントの長さを基準に学習セグメントに置き換え姿勢を復号している．そのため，表 2 における代表姿勢 “89” の生成セグメント（フレーム 2〜29）は，ファイル 19 での身振りとして復元され，発話文字は一部しかマッチしていない．

5.4 トークアニメの生成

生成動作の姿勢パラメータ列とそれに対応する合成音声が得られている．次の課題は，生成動作を映像化し合成音声を当てはめることである．映像化は次のように行う．

(1) 登場キャラクタのモデルを作成，配置する．
(2) それぞれのキャラクタの動作データを読み込む．
(3) 発言の順番に従って，各モデルの動作開始フレームを設定する．
(4) カメラを作成，配置し，有効期間やカメラアングル，カメラの動きを設定する．
(5) 表示画面をモニター画像のみに切り替え，動画を保存する．

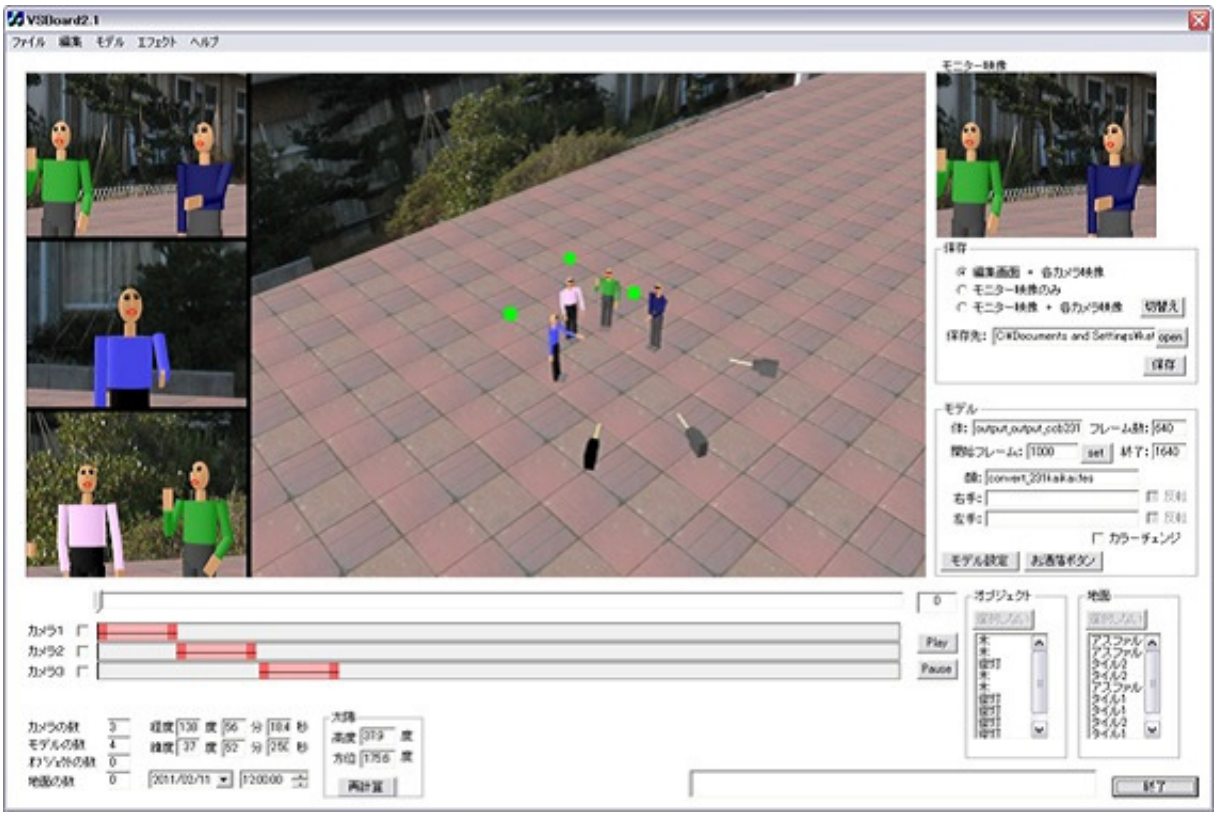


図 6 仮想ロケーションハンティングシステムによるキャスティングとカメラワークの設定
Fig. 6 Casting characters and setting camerawork based on the virtual location hunting system.

この映像化のための編集ソフトとして，文献 [16] で開発した仮想ロケハンシステム VSBoard を使用した．VSBoard は OpenGL をベースに構成されている．VSBoard の編集画面を**図 6** に示す．

この画面中央に舞台構成が示されている．登場キャラクタ 4 名とカメラ 3 台が配置されている．緑色の円はカメラの注視点である．それぞれのカメラから見た映像が，左の小さな画面に示されている．下部のタイムライン上には映像の継続時間が記され，ここでは，カメラ 1，2，3 の順に切り替わる．それぞれのキャラクタには，生成動作が関連付けられ，カメラワークによる最終映像が右上の小さなモニターに表示される．モニター映像は 1 フレームずつの連番画像ファイルとして記録され，さらに動画ファイルとして保存される．作成した映像に，動画編集ソフト AviUtl [1] を利用して，合成音声を加える．

なお，各キャラクタには顔表情が加えられている．顔表情のアニメーション表示手法は数多く提案されている．本研究は身振り手振りを加えた発話のアニメ化が目的なので，顔表情アニメには最低限の動きを加えている．すなわち，実際の顔の表情を FaceAPI [6] で測定し，得られた表情をそのままキャラクタの顔に再現している．そのため，キャラクタの口の動きが発話と一致しないこともある．制作したアニメ映像を Web 上 [17] に公開しておく．

キャラクタの身振りには，当初の意図どおりに芸能人の大きな身振りが表れている．しかし，多くの演技者の身振りを学習しているため，個性が埋没しているようにも見える．キャラクタごとに，特定の演技者の身振りを学習した方が，個性的なキャラクタを構成できると思われる．

また，今回は聴取状態の身振りが考慮されていないので，対話が不自然に見える．無音時の動作も学習しているので，発話していないときの動作も生成可能であるが，無音出力が継続するので単調な動作となろう．相手の発話を出力として聴取状態の身振りを学習すれば，自然な対話が実現できよう．これらは今後の課題としたい．

5.5 上位概念による置き換え

表 2 の発話文字の中には太字がある．これらは，ディスカウンティングにより，状態から強制的に発話された文字である．太字を「・」に置き換え，発話文字の重複を省けば，「海鮮居酒屋　はなの舞はどう　駅前だし」に対する動作は「かい・んい・かや　・な・まい・どう　・・ま・だし」から生成したことになる．居酒屋の店名「はなの舞」は「・な・まい」といった断片的な発話文字を手がかりに動作が生成されている．もし，「はなのまい」という発話が学習データに存在すれば，生成動作は実動作を当てはめることができる．しかし，居酒屋の店名は無数にあり，これらの

すべての店名の発話を学習データに加えることは難しい．

一方，居酒屋は気軽にお酒が飲める店であり，発話者はそのことを強調するために店名によらず同じ身振りをしているかもしれない．また，聞き手にとっては，動作は発話の意味を理解する符丁となりうる．したがって，店名に依存しないしぐさを生成するとよい．

そこで，発話中の言葉が同じカテゴリーに属するならば，同じような動作が生成できるようにする．たとえば，ログの会話中に，「はなの舞」以外にも「壱勢」，「大助」，「葱ぼうず」，「安兵衛」，「船栄」，「ちょこざいや」など多くの居酒屋の店名が見られる．また，時刻を表す言葉として，「何時」，「23:00」，「0 時」などがある．居酒屋の名前が発話されているところは「いざかや」という言葉で，時刻が発話されているところは「なんじ」という言葉で置き換えておく．

あらためてビデオクリップの中から，杯を口にする動作を測定し，これに「いざかや」という発話文字を対応させておく．同様に，指で左腕の時計を指している動作を測定し，これに「なんじ」という発話文字を対応させておく．これらの新たな動作と発話の対を学習データに加える．

これにより，動作を生成すると，これまで居酒屋の店名によって動作が異なっていたのが，同じ動作に統一される．時刻に関しても同様である．アニメの作成には，置き換え前の言葉で発話した音声を当てはめる．実際，書き込み 231 の台詞「海鮮だと大助（いざかや）はお勧めだけど 23:00（なんじ）あと葱ぼうず（いざかや）0 時（なんじ）だったかな」中の居酒屋の名前と時刻を置き換えて，アニメを作成した結果を示す．「いざかや」は杯を口にする動作に，時刻は腕時計を指す動作に置き換わっている．ただし，置き換えられた言葉は，置き換えた言葉と発話の長さが異なるので，動作と多少ずれることもある．制作した比較アニメ映像も Web 上 [17] に公開しておく．

上位概念への置き換えにより，学習する言葉の数を，居酒屋名の例では 7 分の 1 に，時刻の例では 3 分の 1 に節約することができた．今回の実験ではシソーラスは使用せず，居酒屋名や時刻を表す言葉のカテゴリ化と上位概念への置き換えは手動で行った．この置き換えにより，発話が異なっていても意味が同じであれば，同じ動作を生成できる．代表的な動作による置き換えてあるので，聞き手にとって了解しやすい動作と思われる．

一方，同じ言葉でも文脈によって意味が異なる場合には，それぞれ異なる動作を対応付ける必要がある．HMM は文脈に依存しているとはいえ，本手法で意味まで抽出できているかは不明である．言葉の意味の抽出には，発話文字よりも語や句といった大きな単位で台詞を分析する必要がある．自然言語の意味抽出手法の援用により，より汎用的なアニメ製作法への展開が期待できる．

6. おわりに

ビデオ映像中の人物の身振りと発話を抽出し，身振りを状態，発話を出力として両者の関係を HMM により学習した．これにより，台詞を発話するときの身振りを HMM から状態遷移として生成可能となった．このとき，台詞中の言葉を上位概念で置き換えることにより，学習データの貧弱さを補うことができる．電子掲示板のログを対話化し，提案手法によりトークアニメを作成した．

今回は，音声信号から発話文字列への変換，掲示板ログの対話シナリオ化，言葉の上位概念への置き換え，等はすべて手動で行った．将来的には，音声認識や自然言語処理などの手法を援用し，掲示板ログからトークアニメを自動的に生成できることが望ましい．

謝辞 本研究は，一部 JSPS 科研費（26330190）の助成を受けた．

参考文献

[1] AviUtl, available from ⟨http://spring-fragrance.mints.ne.jp/aviutl/⟩.
[2] Brand, M.: Voice puppetry, *Proc. ACM SIGGRAPH*, pp.21–28 (1999).
[3] Bregler, C., Covell, M. and Slaney, M.: Video rewrite: Driving visual speech with audio, *Proc. ACM SIGGRAPH*, pp.353–360 (1997).
[4] Cassell, J., Vilhjálmsson, H.H. and Bickmore, T.: Beat: The behavior expression animation toolkit, *Proc. ACM SIGGRAPH*, pp.477–486 (2001).
[5] CeVIO, available from ⟨http://cevio.jp/⟩.
[6] faceAPI：クレアクト・インターナショナル，入手先 ⟨http://www.creact.co.jp/⟩.
[7] Levine, S, Theobalt, C. and Koltun, V.: Real-time prosodydriven synthesis of body language, *ACM Trans. Graph.*, Vol.28, No.5, pp.1–10 (2009).
[8] Marsella, S., Xu, Y., Lhommet, M., et al.: Virtual character performance from speech, *Proc. ACM SCA*, pp.25–35 (2013).
[9] McNeill, D.: *Hand and Mind: What Gestures Reveal About Thought*, University of Chicago Press (1992).
[10] Neff, M., Kipp, M., Albrecht, I., et al.: Gesture modeling and animation based on a probabilistic re-creation of speaker style, *ACM Trans. Graph.*, Vol.27, No.1, Article 5 (2008).
[11] Stone, M., Decarlo, D., Oh, I., et al.: Speaking with hands: creating animated conversational characters from recordings of human performance, *Proc. ACM SIGGRAPH*, pp.506–513 (2004).
[12] Sun, L., Shoulson, A., Huang, P., et al.: Animating synthetic dyadic conversations with variations based on context and agent attributes, *Comp. Anim. Virtual Worlds*, Vol.23, pp.17–32 (2012).
[13] VoiceTEXT, available from ⟨http://voicetext.jp/⟩.
[14] 高木幹夫，下田陽久監修：新編　画像解析ハンドブック，東大出版会 (2004).
[15] 山本正信：ドリフト修正機能を有する動画像からの身体動作推定法，電子情報通信学会論文誌，Vol.J88-D-II, No.7, pp.1153–1165 (2005).
[16] 山本　遼，遠藤大輔，菅原彩子ほか：仮想現実空間による

映画制作のロケーションハンティング，映像情報メディア学会誌，Vol.62, No.6, pp.889–900 (2008).
[17] 山本研究室 HP，入手先 ⟨http://www.vision.ie.niigata-u.ac.jp/motion-capture-t_animation.html⟩.

付　　録

A.1 ビデオクリップ

学習データとして使用したビデオクリップで発話された台詞を示す．なかには重複も見られる．ファイル 1 はファイル 5 の一部であり，ファイル 9 とファイル 15 は同じビデオクリップがソースであるが，聞き取りにくかったので 2 通りの台詞として収録した．

1 ダイノジ 2-1　39　アンジェラアキだね
2 ダイノジ 5　70　きょうもまた俺がばんばんぼけんだからよ　びしばし突っ込み
3 キャンキャン　109　ふーるいアルバムの野郎，めくっちまったらサンキュって呟いたってんだ
4 ダイノジ 2-2　91　読書の秋とか，そういうのでお前，こっちはなアンジェラアキだねってぼけてんだ
5 ダイノジ 1　99　いいですか，みなさん．パクリはだめですからね，ほんとすいませんね．あーい
6 ダイノジ 4　53　誰がそのまんまなんだよ，お前よ
7 フットボールアワー　59　ふつうにうまいだけやったん．あまいか
8 フルーツポンチ　74　ごめんごめん，ちょっとまって．今雲がすごい
9 ギャロップ 1-1　39　ちゃうやん．なんでそんな頭なってん．
10 スマイル 2　40　ごめんごめん．もう言わん言わへん
11 ジャルジャル　41　お前がやってつってん，わろて
12 ギャロップ 3　48　関係あらへん．
13 スマイル 1　77　すいませんねえ　こっちはこっちで
14 ギャロップ 1-1　57　ちが　なんでそんな頭なったんや
15 ギャロップ 1-2　58　なんでそんな頭なったんやいうてんねん
16 ノブシコブシ　53　そんな状態で本当に大丈夫なん
17 NON STYLE1　115　ちょっと僕ね，この世に蔓延るいじめっていうのをおれの力で止めたいねん
18 NON STYLE2　64　俺らはハートフル漫才師や
19 NON STYLE3 58　俺がボスみたいなるやろ
20 NON STYLE4　74　仲良しやないか
21 ギャロップ 2　47　お前が抜けすぎ
22 女と男 1　38　女です　よろしくお願い
23 女と男 2　72　ぶりっこでなあ　自己紹介が　ごっつ腹立ってん
24 女と男 3　43　むにっ．
25 女と男 4　76　このやろ料理している後ろから，ボブスレーで突っ込んだろか
26 スマイル 3　87　携帯変えた　しまっていこうぜ
27 しずる 2　94　歯磨いた後に，これ食った所でまずいだけ
28 しずる 3　21　自分
29 ダイノジ 3　37　そこつっこむんだろ，おめえはよ
30 スマイル 4　73　いやそっちのセリフやん
31 TKO　29　暴力反対

A.2 掲示板ログ

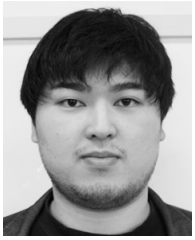
226 名前：雪ん子 [] 投稿日：2008/01/18(金) 21:40:33 ID:****** [******]
何度か書き込みしてますが・・・・
新潟市内で海鮮メインで夜遅くまで営業してる居酒屋教えてください＾＾v
29 日火曜に行きます！！！！
227 名前：雪ん子 [] 投稿日：2008/01/18(金) 22:48:52 ID:****** [******]
>>226
海鮮居酒屋　はなの舞はどう？　駅前だし
http://r.gnavi.co.jp/r017100/
228 名前：雪ん子 [sage] 投稿日：2008/01/18(金) 23:47:31 ID:****** [******]
壱勢とか．
229 名前：雪ん子 [] 投稿日：2008/01/19(土) 00:22:19 ID:****** [******]
227 さんへ
はなの舞ってチェーン店居酒屋ですよね；；；
せっかくの情報ですが　　わざわざ新潟まで行くので地元で有名オススメ居酒屋
を知りたいのです＾＾；　　私の聞き方悪くてごめんなさい；；
230 名前：雪ん子 [sage] 投稿日：2008/01/19(土) 00:43:46 ID:****** [******]
>>229
>>228
231 名前：雪ん子 [sage] 投稿日：2008/01/19(土) 11:36:15ID:****** [******]
>>229
遅くまでって何時頃？新潟は早じまいが多いよ．
海鮮だと大助はお勧めだけど 23:00、あと葱ぼうず、0 時だったかな．
新潟のグループでは安兵衛、船栄、ちょこざいや系とかがメジャー．
後はホットペッパー検索しておすすめレポートなど参考にした方が吉．
店の雰囲気も分るしお勧め．

加藤 和輝

2009 年新潟大学工学部情報工学科卒業．2011 年同大学院自然科学研究科情報理工学専攻博士前期課程修了．在学中はパターン認識，コンピュータアニメーション等の研究に従事．

今村 友哉

2014 年新潟大学工学部情報工学科卒業．同年株式会社ジャパンネット入社．在学中は，コンピュータアニメーションの研究に従事．



石本 貴康

2008 年新潟大学工学部情報工学科卒業．同年株式会社プレスメディア入社．在学中はコンピュータアニメーションの研究に従事．

渡部 郁美

2008 年新潟大学工学部情報工学科卒業．同年セイコーエプソン株式会社入社．在学中は，コンピュータアニメーションの研究に従事．

山本 正信 （正会員）

1973 年九州工業大学工学部制御工学科卒業．1975 年東京工業大学大学院理工学研究科制御工学専攻修士課程修了．同年電子技術総合研究所（現，産業技術総合研究所）入所．モーションキャプチャ，コンピュータアニメーション等の研究に従事．1989～1990 年カナダ国立研究協議会招聘研究員．現在，新潟大学工学部情報工学科教授．工学博士．電子情報通信学会フェロー，IEEE，ACM 各会員．